Skin Disease Basic Analysis

A case study for a computational biology framework

T.P.A. Beishuizen (0791613)
Biomedical Engineering - Computational Biology
Computer Science - Data Mining
Eindhoven, University of Technology
Email: `t.p.a.beishuizen@student.tue.nl`

March 14, 2018

Contents

1 Introduction

Biomedical Engineers try to extract useful information out of biomedical data. Biomedical data is a generalizing term that describes multiple data types[1]. Examples of biomedical data are microarray data[2], mass spectrometry datas[3, 4] and nuclear magnetic resonance data[5], but also clinically derived data[6, 7]. From a bioinformatics perspective these biomedical data types vary significantly[1] and therefore extracting information out of biomedical data is not a trivial task. A framework for biomedical data analysis can help guiding biomedical engineers in their process of information extraction from their biomedical datasets. The framework can provide different options in processing the data, taking into account common dataset issues[8, 9, 10] and approaches to reach a certain goal[11, 12]. Such frameworks are proposed and discussed, however mainly focus on the integration of databases[13, 14], are made specifically for one research area[15, 16, 17] or is limited to one specific type of analysis[18]. A framework that combines database integration, multiple research areas and multiple types of data analysis would be very beneficial for biomedical engineers, guiding them through their biomedical data analysis projects.

To find out what techniques and guidelines are useful to put in such a framework, a case study is done with an example biomedical dataset. This example dataset is a based on gene expression of skin diseases[19, 20, 21, 22, 23, 24, 25, 26] and is a microarray dataset. Two skin diseases were intended to be examined with this dataset, psoriasis and atopic dermatitis. The expression for a total of 54675 genes were measured for skin disease patients on skin affected by the disease (lesional skin) and skin not affected by the disease (non-lesional skin). Further data from healthy subjects was also acquired. Nine datasets were available, six for psoriasis[19, 20, 21, 22, 23] and three for atopic dermatitis[24, 25, 26]. The number of tested skin biopies ranged from 28 to 180 per dataset whereas the number of measured genes is the same for every set, namely 54675.

In this project, a basic analysis was done on the available skin disease datasets, as a case study for creating a framework for biomedical engineers. First, a background is given on the dataset. Secondly, methods to extract information from the data are explained, followed by their results. At last, the results are discussed and useful aspects for a biomedical engineering framework about the basic analysis are concluded from the case study.

2 Skin Diseases Datasets

Skin diseases can have a major impact in someone's life. Whereas skin diseases are not as life threatening as diseases such as cancer, Alzheimer and AIDS, they can lower quality of life significantly. When looking at the health-related quality of life (HRQL), patients with psoriasis show the same problems as patients with other major chronic health conditions[27]. Patients with psoriasis or atopic dermatitis suffer from severe itching and pains. Further insights into the skin diseases can help alleviate their unwanted side-effects and help improve the patients' quality of life[28].

Information on both of these skin diseases can be found in nine datasets stored on the NCBI database[29]. The datasets comprise microarray data extracted biopsies of psoriasis patients, both from their lesional and non-lesional skin. In several experiments this skin is taken from the same patient. Also some skin is taken from patients not suffering from the diseases at all. six datasets gathered data from Psoriasis patients and three from atopic dermatitis patients. These datasets consist of a total number of 54675 genes, the features of the dataset. The range of acquired samples varies among datasets from 28 tot 180. Also, since every dataset was created by different people, some minor differences can be present in them as well (Table 1), due to different measurement equipment.

The nine datasets are rich in information. The dimensionality is very high and if combined the datasets also have a decent number of samples, from a data mining perspective. Several challenges arise in the dataset, too, as in biomedical datasets often have. Here, three of these challenges are discussed.

At first the challenge of handling nine different datasets is essential. Even though the genes were chosen according to the Affymetrix Human Genome[30], the layouts are not identical. These

Table 1: Details of the nine skin disease datasets. The number of samples and genes has been given, as well as remarks of the skin types.

Disease	Dataset name	Sample size	Genes	Remarks
Psoriasis	**GSE13355** [19]	180	54675	Three skin types: - NN (normal, 64 samples) - PN (non-lesional, 58 samples) - PP (lesional, 58 samples)
	GSE30999 [20]	170	54675	- No normal patients - Non-lesional (85 samples) - Lesional (85 samples)
	GSE34248 [21]	28	54675	- No normal patients - Non-lesional (14 samples) - Lesional (14 samples)
	GSE41662 [21]	48	54675	- No normal patients - Non-lesional (24 samples) - Lesional (24 samples)
	GSE78097 [22]	33	54675	Different types of skin samples: - Normal (6 samples) - Mild Psoriasis (14 samples) - Severe Psoriasis (13 samples)
	GSE14905 [23]	82	54675	- Normal skin (21 samples), - Non-lesional skin (28 samples) - Lesional skin (33 samples)
Atopic Dermatitis	**GSE32924** [24]	33	54675	- Normal skin (8 samples) - Non-lesional skin (12 samples) - Lesional skin (13 samples)
	GSE27887 [25]	35	54675	Different type of skin samples, pre and post treatment of skin: - Pre non-lesional (8 samples) - Post non-lesional (9 samples) - Pre lesional (9 samples) - Post lesional (9 samples)
	GSE36842 [26]	39	54675	Also difference between acute and chronic dermatitis. - Normal (15 samples) - Non-lesional (8 samples) - Acute lesional (8 samples) - Chronic lesional (8 samples)

difference originate from the intended research goals and the data availability. It is not possible to just concatenate samples without some form of preprocessing. Only the parts that are the same all over the datasets must be taken and all other parts omitted.

A second challenge can be found in the high number of genes. There were 74675 genes measured, averaged at about 1000 times the number of samples. The genes that are significantly involved in the skin diseases is however expected to be much lower than the total number of measured genes. Many genes are redundant and can be removed during preprocessing, a valuable and complex step in biomedical data mining.

The third challenge is about data volume. The number of samples differs from 28 to 180, all of them being a very low number compared with the number of genes. From a data mining perspective, this number of samples is hard to gather information from. This can create problems, mainly during machine learning, with such a low training and test set. Several cases will arise where all training and test set agree with the algorithm, whereas other samples from the sample space would not.

2.1 Additional Data

The genes all correspond to the same genes for all of these nine different datasets. The NCBI database[29] also provides separate data containing substantial information for every gene. This information includes gene ID, commonly known name and abbreviation and which gene database it originates from. It also contains processes and molecular reactions the gene is involved with, as well as the location of it in the cell.. This data can be used to find links between several processes and their corresponding genes.

3 Methods

Three different aspects were investigated with the dataset. At first several techniques were used to reduce the high number of genes. Secondly the genes (either before feature reduction or after) were clustered. Thirdly all genes of psoriasis and atopic dermatitis were compared after reducing them to a lower number of genes, to find out whether genes had been over- or underexpressed in both of them.

A previous project[1] found out that healthy and non-lesional skin do not show many differences and after some testing, the same conclusion was reached in the current study. Aside from this, the difference between lesional and non-lesional skin was the most important for skin diseases to compute, as that difference showed which genes are over- or underexpressed in lesional skin. Therefore the main focus of the project was showing the difference between non-lesional and lesional skin in terms of genes.

For the computation of feature reduction, the largest possible dataset was created with only non-lesional and lesional skin samples. The biggest dataset that could be created was with the Psoriasis sets that had both non-lesional and lesional skin (table 1), good for a total of 423 samples. When comparing Psoriasis and Atopic Dermatitis, all suitable samples of Atopic Dermatitis were collected for a total of 58 samples. In gene reduction and clustering only the psoriasis dataset was used, due to its bigger sample size. When comparing pPsoriasis and atopic dermatitis, both were used.

3.1 Feature Reduction

Since the number of features in the data was 54675 (the number of genes), a significant feature reduction was needed before any meaningful computations could be done. Therefore two different ways of feature reduction were explored. The first one was doing a simple t-test to find all genes that were significantly different. The second one was testing correlation between all genes, also known as multicollinearity testing. The complete layout of the feature reduction was visualized for understanding (Figure 1).

[1] BEP Project - *Manouk Groels*

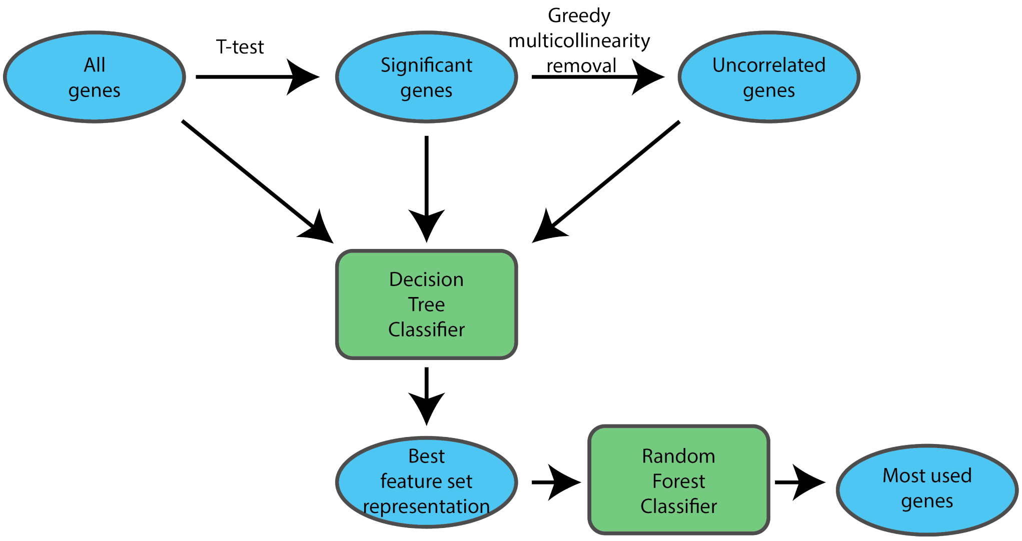


Figure 1: A flowchart layout of the feature reduction. The complete data set first was reduced by using a t-test and next by the greedy clustering algorithm. The three feature set representations were all three tested by a decision tree classifier and the one with the best results was chosen. At last the best feature set representation was classified with a random forest classifier and the genes that were most used in the final random forest were selected.

For using a t-test, the SciPy package was used. Three different t-test possibilities are available, one for paired data, one for data with equal unknown variance and one for data with unequal unknown variance[2]. Two of these t-tests were used, the paired t-test and the t-test for unequal unknown variance. Almost all datasets (GSE13355, GSE30999, GSE34248 and GSE41662) were paired, in those cases the paired t-test was used (Equation 1) and in the remaining set (GSE14905) plus the combined set (all datasets together) the unequal variance t-test were used (Equation 2). In these equations $\bar{y}$ was the mean of a distribution, s the standard deviation, n the number of samples, $\bar{d}$ the average value for difference in paired samples and t was the normalized version of the gene to be checked ($t = \frac{y-\bar{y}}{s/\sqrt{n}}$).

$$\mathbb{P}(t > |t_{calc}|) \text{ with } t_{calc} = \frac{\bar{d}}{s_{\bar{d}}} \tag{1}$$

$$\mathbb{P}(t > |t_{calc}|) \text{ with } t_{calc} = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{2}$$

A further reduction could be done by removing multicollinearity. Multicollinearity means that multiple genes are highly correlated with each other. Usually this would be computed by calculating the correlation coefficient between all genes. This coefficient however was computationally quite heavy to compute and therefore an alternative approach was used (Algorithm 1 and Gigure 2). A greedy hands on approach computed correlation between genes and removed multicollinearity. For this greedy hands on approach a list of genes was used, ordered on the significant difference of gene expression between non-lesional and lesional skin data, computed by the t-test.

[2] Biomedical Data Analysis - *Tim Beishuizen*

Algorithm 1 The greedy multicollinearity removal algorithm

```
procedure GreedyMulticollinearityRemoval($SortedGenes$)
    $UncorrelatedGenes \leftarrow emptyList$
    $correlationValue \leftarrow 0$
    for $newGene$ in $SortedGenes$ do
        for $oldGene$ in $UncorrelatedGenes$ do
            if correlation($oldGene, newGene$) $>$ $correlationValue$ then
                $correlationValue \leftarrow$ correlation($oldGene, newGene$)
            end if
        end for
        if $correlationValue < 0.7$ then
            $UncorrelatedGenes \leftarrow UncorrelatedGenes \cup newGene$
        end if
        $correlation \leftarrow 0$
    end for
    return $UncorrelatedGenes$
end procedure
```

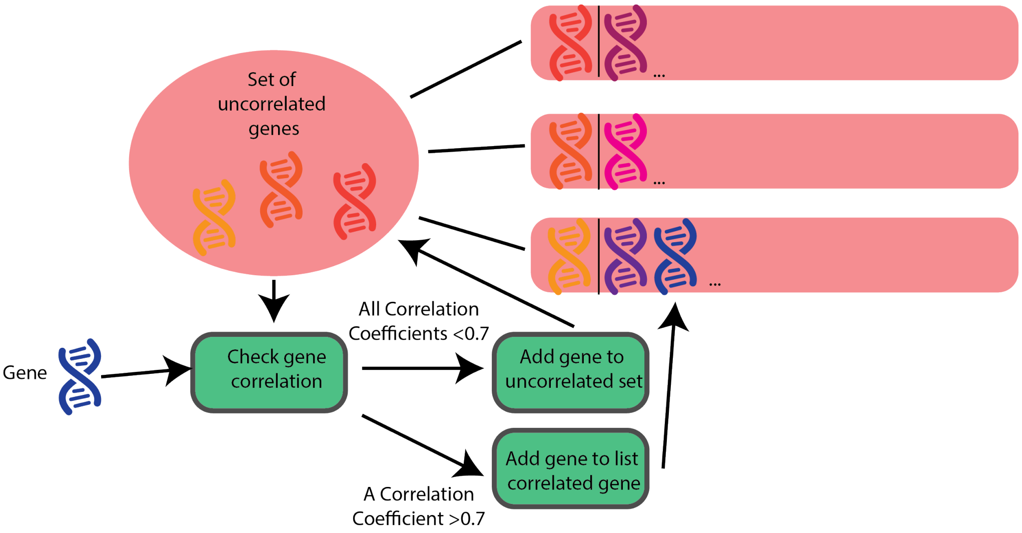
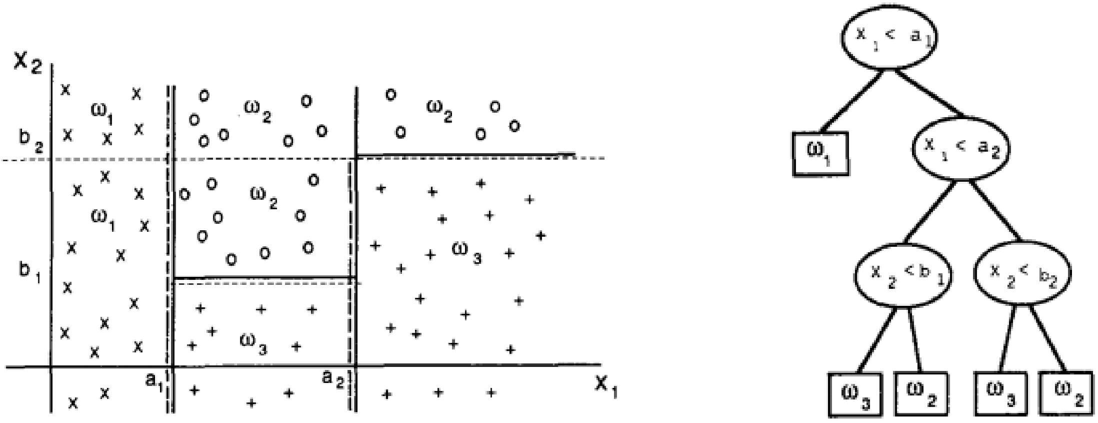

Set of uncorrelated genes
...
...
...
Gene
Check gene correlation
All Correlation Coefficients <0.7
Add gene to uncorrelated set
Add gene to list correlated gene
A Correlation Coefficient >0.7

Figure 2: A flowchart layout how the gene multicollinearity removal algorithm worked. The correlation coefficient is iteratively computed between a new gene and the uncorrelated set of genes. If all correlation coefficients are lower than 0.7, the new gene did not correlate enough with the set of genes and therefore is added to the uncorrelated set of genes. If it had a correlation of higher than 0.7 for another gene, it was added to that the list for the gene it is correlated with.

After removing the unsignificant genes and multicollinearity, a decision tree classifier (Figure 3) from the scikit-learn package was used to find out whether removing unusable genes and multicollinearity actually improved the possibility to better classify lesional skin. All three gene sets (all genes, significant genes and greedy clustered genes) were used for decision tree classification and split in 80% training and 20% test set. A cross validation was done with 100 different subsets for this classification, dividing the 80% training set into 100 subsets that all were used as a validation set once. With the cross validation, both a validation score and a test score were made to evaluate the performance with the three gene datasets.



(a) A two dimensional sample space divided in three regions.

(b) The corresponding decision tree

Figure 3: An example of how a decision tree divides a sample space in different regions. A sample space (Figure 3a) with features (x_1 and x_2) is divided in regions (ω_1, ω_1 and ω_1). A decision tree (Figure 3b) shows how this division is done using thresholds for the features. Every node in the tree corresponds to dividing the sample space in different regions using a threshold for a feature and every leaf corresponds to a region. The classifier tries to divide the sample space in regions finding the best feature threshold combinations, until boundary conditions are met.

After testing with a decision tree classifier a random forest classifier was used. A random forest classifier creates not one, but multiple decision tree classifiers on subsets of the complete sample space and averages the outcome, for this case 1000 decision trees were created. The input of the random forest classifier was the best gene dataset of the three tested gene sets (all genes, significant genes, uncorrelated genes), according to the decision tree classifier. Aside from the input, the decision trees were only allowed to have a depth of two, as only the genes that show the biggest changes are important. Genes that were used multiple times as a splitting criterium in the created decision trees were extracted from the random forest classifiers. These genes occurred most frequently as the best splitting criterium and therefore had the best relation with the skin disease. The six genes that occurred most were selected.

3.2 Clustering

Given the high number of genes even after feature reduction, clustering is a good way to find whether genes show similar behaviour. Two different ways of clustering were explored. The first way is based on the values of the genes, clustering genes that showed the same behaviour. A second way is clustering them based on biomedical relations between the genes. After exploring these two ways, the interesting results were combined. A complete layout of every thing done during this clustering was visualized (Figure 4).

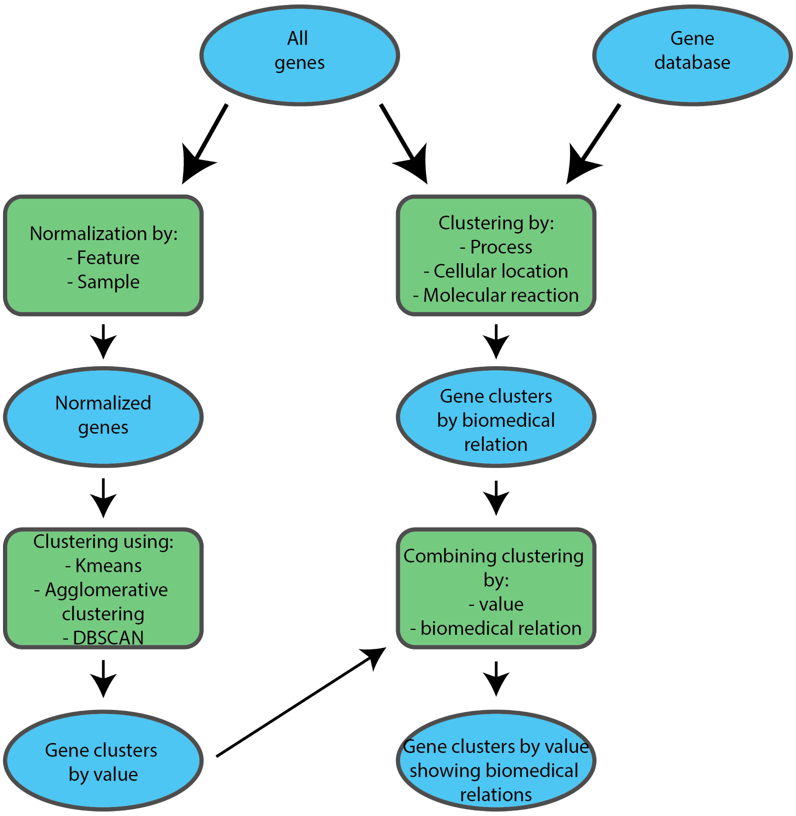


Figure 4: A flowchart layout how clustering was done. Two types of clustering were used, by biomedical relation and by value. For the clustering by value, first a normalization was done. After both types of clustering were done, they were combined and biomedical relations within the clusters were shown.

Before clustering was done, the values were normalized. This normalisation was done to remove high differences in especially variation. Two types of normalization were done (Figure 5). At first normalization per genes was done, so every gene was treated equally. Relatively big differences in gene expression were observed that way. Normalization per sample gives skin cell specific genes higher values than non-skin cell specific genes, because their expression should be higher overall. This way genes known to be more prevalent in skin cells had a higher chance to be noticed for the difference between lesional and non-lesional skin.

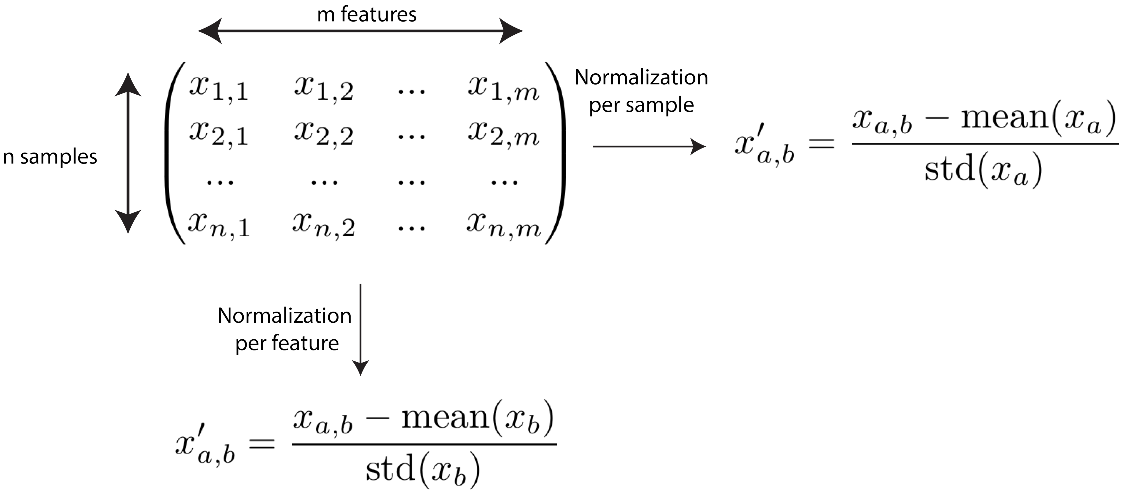

m features
n samples
$x_{1,1}$ $x_{1,2}$ $\cdots$ $x_{1,m}$
$x_{2,1}$ $x_{2,2}$ $\cdots$ $x_{2,m}$
$\cdots$ $\cdots$ $\cdots$ $\cdots$
$x_{n,1}$ $x_{n,2}$ $\cdots$ $x_{n,m}$
Normalization per sample
$x'_{a,b} = \frac{x_{a,b} - \text{mean}(x_a)}{\text{std}(x_a)}$
Normalization per feature
$x'_{a,b} = \frac{x_{a,b} - \text{mean}(x_b)}{\text{std}(x_b)}$

Figure 5: An explanation of the difference between normalization per feature and normalization per sample. $x_{a,b}$ is value x for sample a and feature b, mean() gives the mean of the values given (either a sample row or feature column) and std() gives the standard deviation for the values given. For both the formula of how to compute the normalized value $x'_{a,b}$ is given.

The first type of clustering was done by using basic clustering methods. Three different types of clustering were used (Table 2): K-means, Agglomerative and DBSCAN, all three of them methods in scikit-learn. For both K-means and agglomerative clustering a variation of cluster numbers is chosen to find the best possible selection.

Table 2: The three different clustering types for clustering by value.

Clustering type	**Explanation and parameters (if applicable)**
K-means	Assigns genes to clusters closest to cluster centers Changes cluster centers after new genes are assigned Tested number of clusters: 5, 8, 9 10, 11, 12, 15, 20.
Agglomerative (Hierarchical)	Adds genes closest to each other together in a cluster until only the desired number of clusters are left. Tested number of clusters: 5, 8, 9 10, 11, 12, 15, 20
DBSCAN	Clusters gene groups with high density together. Good with data with high number of features and low number of samples

Aside from clustering the genes by values, another way was clustering by biomedical relations. The additional dataset (Subsection 2.1) links genes to three different biomedical components (Table 3): processes, cellular locations and molecular reactions. Genes linked to the same biomedical relation means the genes are involved in that particular process and therefore can be clustered together. Since genes can be linked to multiple processes, locations and reactions, they were put in multiple clusters for all three.

Table 3: The three biomedical relation types available in the database (subsection 2.1) with a description and several examples.

Biomedical relation	**Explanation**	**Examples for gene:** *discoidin domain receptor tyrosine kinase 1*
Process	Biomedical processes the gene product is involved with	- regulation of cell growth - peptidyl-tyrosine autophosphorylation - lactation
Cellular location	The gene product is located in or part of certain cellular locations	- extracellular region - plasma membrane - extracellular vesicular exosome
Molecular reaction	A molecular reaction the gene product is involved with	- nucleotide binding - collagen binding - protein tyrosine kinase collagen receptor activity

Both types of clustering should give different results, as one of them only took into account the values and while the other focused on biomedical relations. A good way to find results was to combine both clustering types, to find out which value clusters were also involved in similar biomedical processes and vice versa. This was done in two different ways. The first way was done by ordering the biomedical relation dependent clusters by their difference between lesional and non-lesional gene expression. More difference between these expressions means it is regarded as showing a bigger difference between lesional and non-lesional skin.

The second one was by selecting all value dependent clusters. For all of these clusters biomedical relations were investigated. If clusters showed that multiple genes in that cluster were related to an interesting process, cellular location or molecular reaction, the cluster would be taken as more interesting.

At last it is important how to visualize the clusters. Every gene has 423 different values, one for every sample. this however is hard to visualize. Therefore, the gene values were averaged in two different values, one value corresponding to lesional skin and one to non-lesional skin. Gene plots were made with every gene being a data point with the x-value being the non-lesional skin gene expression and for the y-value the average lesional skin gene expression.

3.3 Psoriasis Versus Atopic Dermatitis

At last a brief search for the difference between Psoriasis and Atopic Dermatitis is done. Since up- and down regulations for both of these diseases are more interesting, first feature reduction is done by only choosing the ones that showed a significant difference. The genes for both diseases are then compared and for all genes that are up- or down regulated for both diseases, all biomedical processes, cellular locations and molecular reactions are extracted and counted how many times they were present for both diseases. The five most occurring processes, cellular locations and molecular reactions were given, if interesting enough.

4 Results

The results were presented in the same order as the methods. First, the results for feature reduction were shown for the t-test, multicollinearity reduction as well as for the machine learning improvements. Then, the results for clustering were shown. Emphasis was put on combining clustering by value and clustering by biomedical relations. Finally, the similarities between Psoriasis and Atopic Dermatitis were shown to see if genes, processes, cellular locations and molecular reactions are involved in both diseases.

4.1 Feature Reduction

The t-test was used between non-lesional and lesional skin for every dataset separately and combined (Table 4). A low p-value of $p = 0.001$ was chosen by trial and error. Interestingly enough when all datasets were combined, the number of significant features was lower than when looking at the datasets separately. This could indicate that dataset specific noise was present in the datasets and heterogeneity actually improved noise reduction. After this feature reduction, further computation was done with the 1768 genes left.

Table 4: The results of using the t-test for genes in all relevant psoriasis datasets separately and combined. A paired t-test was done if applicable, otherwise an unknown variance unpaired t-test was done. The number of samples for both lesional and non-lesional skin were shown additionally.

Dataset	**Samples Lesional Skin**	**Samples Non-Lesional Skin**	**Paired**	**All Genes**	**Significant Genes**
GSE13355	58	58	Yes	54675	22106
GSE30999	85	85	Yes	54675	20836
GSE34248	14	14	Yes	54675	7824
GSE41662	24	24	Yes	54675	15672
GSE14905	33	28	No	54675	13355
Combined	214	209	No	54675	1768

Computing multicollinearity was done next. Greedy clustering yielded 335 clusters for the combined dataset, so for 335 genes the correlation coefficient was lower than 0.7 between all of them. The ability to find the difference between lesional and non lesional skin with a decision tree classifier showed that only using significant genes gave the best results, whereas using the uncorrelated genes made both the validation and test score worse. (Table 5).

Table 5: The results of the decision tree classifier for different sets of genes: all genes, only the significant genes and the uncorrelated genes. The decision tree was cross validated by division in 100 different subsets and afterwards tested by a separate test set.

Gene set	**Genes**	**Validation score**	**Test score**
All genes	54675	0.584	0.575
Significant genes	1768	0.959	0.943
Uncorrelated genes	335	0.928	0.915

With a validation and test score of around 0.95, the significant gene set was used for a random forest classification consisting of 1000 decision trees. The genes that were used as splitting criteria for these decision trees were collected to find out whether the same genes pop up multiple times as the best splitting criteria (Table 6). While the number of multiple occurrences was not as high as expected, the six most occurring genes already showed to have a relation with skin diseases.

Table 6: The six most used genes as splitting criteria for the random forest classifier. The genes were also compared with literature.

Gene ID	Gene Title	Times used in random forest	Link with Psoriasis
NM_004262	transmembrane protease, serine 11D	45	Previously discovered relation with Psoriasis[20]
AI186548	keratin 77	33	Keratin is upregulated in uncontrollable growth skin cells[20]
BF032500	MACRO domain containing 2	33	Previously discovered relation with Psoriasis[20]
NM_001062	transcobalamin I (vitamin B12 binding protein, R binder family)	31	Previously discovered relation with Psoriasis[20]
NM_005621	S100 calcium binding protein A12	31	Previously discovered relation with Psoriasis[20]
U19557	serpin peptidase inhibitor, clade B (ovalbumin), member 3 & 4	31	Previously discovered relation with Psoriasis[20]

4.2 Clustering

Only the 1768 significant genes found from feature reduction were used for clustering. The genes were clustered per process, cellular location and molecular reaction they were linked to, so if two genes were both related to a certain process, they were clustered together. Since a gene could be linked to multiple processes, it could also be present in multiple clusters. The sixteen highest scoring processes (Appendix A and Figure 6), cellular locations (Appendix A) and molecular reactions (Appendix A) were shown for scaling per gene and scaling per sample. The findings on the cellular locations and molecular reactions of the genes did not show any interesting results. The interesting results were found in several interesting processes, related to inflammatory response. All of these showed up regulation in the sample and were plotted to show their gene locations (Figure 6)

- acute inflammatory response
- response to interferon gamma
- chronic inflammatory response
- oxidative stress.

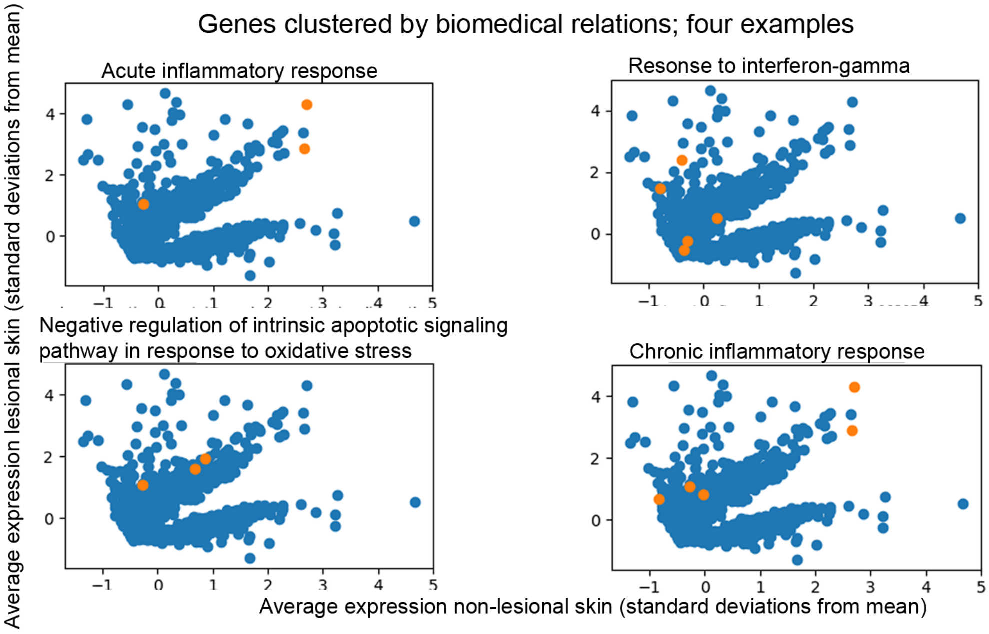


Figure 6: Four examples of processes that showed a high difference in gene expression between lesional and non-lesional data. The data was normalized per gene and orange points are genes that were related to the process, blue points were genes unrelated to the process

Cross-referencing the value-based clustering and the biomedical relation clustering, agglomerative clustering with ten clusters gave the best results (Figure 7). After linking the genes with biochemical processes, cluster 5 and cluster 8 showed interesting links in results. Cluster 5 was linked to multiple processes typically occurring in skin cells, for example involving keratin and the epidermis. Cluster 8 was linked to multiple processes related to negative regulation of several enzymes, for example endopeptidase, peptidase and proteolysis. Other clusters were linked to less specific processes, e.g. transport and protein binding.

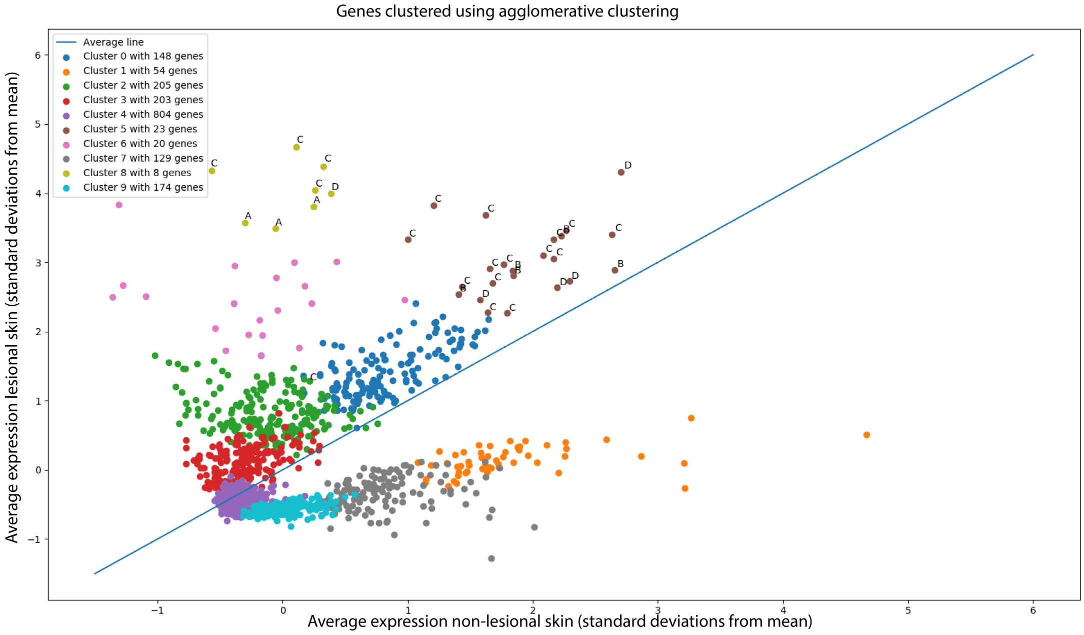


Figure 7: 10 different clusters found by agglomerative clustering for the Psoriasis dataset. The data points were genes that showed a significant difference in expression between lesional (y-axis) and non-lesional (x-axis) skin and normalized per sample. Cluster 5 and 8 showed the most useful genes and its genes were therefore also marked for Psoriasis relevance: Known Psoriasis marker (A), Known to be upregulated for Psoriasis (B), Known to be upregulated for uncontrollable growth (C) and unknown relation (D)

Since cluster 5 and 8 gave interesting results on a process level, literature was used to link them to psoriasis as much as possible (Appendix B). Most of the genes in these clusters were linked to psoriasis directly or indirectly (Figure 7).

4.3 Psoriasis Versus Atopic Dermatitis

At last the significant genes for atopic dermatitis were compared with the significant genes of psoriasis (Table 7). After matching both datasets to find out whether any genes were significant in both datasets, 96 genes were left. These genes were linked to processes, cellular locations and molecular relations to find biomedical relations that showed differences between lesional and non-lesional skin in both psoriasis and atopic dermatitis. Whereas the results from cellular locations and molecular reactions were inconclusive, the processes gave some more results. Of these, five processes were related to biggest number of genes. All of these processes, however, seemed fundamental and therefore not useful for further investigation:

- metabolic process
- small molecule metabolic process
- regulation of transcription, DNA-templated
- signal transduction
- transcription, DNA-templated

Table 7: The initial conditions of the Psoriasis and Atopic Dermatitis datasets

Dataset	Samples Lesional Skin	Samples Non-Lesional Skin	All Genes	Significant Genes
Psoriasis	214	209	54675	1768
Atopic Dermatitis	30	28	54675	516

5 Discussion

Three different aspects have been researched: feature reduction, clustering and comparing Psoriasis and Atopic Dermatitis. Some results do not give any additional insights, however others actually show possible future research topics.

Using a t-test seemed to be an efficient way to reduce the number of genes significantly without any loss in information. Especially for multiple and bigger datasets this method was effective for Psoriasis. The loss of information was present however when using the greedy correlation method, which was not used in further computations because of that information loss. After computing which genes would be used as splitting criteria, the six most occurring genes all but one either can be found in literature having a relation with Psoriasis. The remaining gene would be logical to be related to Psoriasis as well and therefore the significant gene set seems a good set to continue with.

Initially, clustering by biomedical attribute and clustering by value did not give good results. No genes seemed interesting enough by using only one type of clustering. However when combining both of them, the genes that were clustered together by both values and biomedical process gave two clusters of genes that showed to have a relation with both each other as well as with psoriasis. After labelling these genes multiple are already linked with psoriasis directly or indirectly, however some would be interesting to add to the linked genes.

6 Conclusion for Framework

The skin disease datasets were examined as a case study which aspects are important for a biomedical framework. This case study lead to three different insights that would be helpful to add to a framework for general research: Global analysis, feature dimensionality reduction and database integration

Initially, a global analysis of the data set would be beneficial at the start of a project. The goal of this global analysis would mainly be to understand the how the data looks like from the outside. Means and variances should be shown for both samples as features and irregularities such as missing values and outliers should be made visible, so the one using the dataset, understand how it works. Another important part for global analysis would be the multicollinearity, whether features or samples are very similar and therefore could cause problems for the user if not knowing these similarities. A global analysis would help understanding the data quicker and therefore create higher efficiency early on.

On the subject of feature dimensionality reduction some simple approaches have been used. Using a t-test showed quite a good result, whereas multicollinearity testing lacked a possibility to do that for high number of features. Some initial insights were obtained what to add and in which directions more research should be done to add more strategies to cope with feature dimensionality.

A last topic that would be useful for the framework is something not thought of before. Information about known biomedical relations between genes, metabolites or other biomedical substances is very valuable. This information would be very helpful if possible to automatically use it during the research. An extension that could do that is therefore something to consider during the framework creation.

References

[1] N. Gehlenborg, S. I. O'donoghue, N. S. Baliga, A. Goesmann, M. A. Hibbs, H. Kitano, O. Kohlbacher, H. Neuweger, R. Schneider, D. Tenenbaum, *et al.*, "Visualization of omics data for systems biology," *Nature methods*, vol. 7, no. 3s, p. S56, 2010.

[2] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, *et al.*, "Minimum information about a microarray experiment (miame)—toward standards for microarray data," *Nature genetics*, vol. 29, no. 4, p. 365, 2001.

[3] J. S. Cottrell and U. London, "Probability-based protein identification by searching sequence databases using mass spectrometry data," *electrophoresis*, vol. 20, no. 18, pp. 3551–3567, 1999.

[4] K. Dettmer, P. A. Aronov, and B. D. Hammock, "Mass spectrometry-based metabolomics," *Mass spectrometry reviews*, vol. 26, no. 1, pp. 51–78, 2007.

[5] D. Capitani, A. P. Sobolev, and L. Mannina, "Nuclear magnetic resonance–metabolomics," *Food Authentication: Management, Analysis and Regulation*, p. 177, 2017.

[6] B. Liu, X. Zhou, Y. Wang, J. Hu, L. He, R. Zhang, S. Chen, and Y. Guo, "Data processing and analysis in real-world traditional chinese medicine clinical data: challenges and approaches," *Statistics in medicine*, vol. 31, no. 7, pp. 653–660, 2012.

[7] D. F. Sittig, A. Wright, J. A. Osheroff, B. Middleton, J. M. Teich, J. S. Ash, E. Campbell, and D. W. Bates, "Grand challenges in clinical decision support," *Journal of biomedical informatics*, vol. 41, no. 2, pp. 387–392, 2008.

[8] P. Bertolazzi, G. Felici, P. Festa, and G. Lancia, "Logic classification and feature selection for biomedical data," *Computers & Mathematics with Applications*, vol. 55, no. 5, pp. 889–899, 2008.

[9] G. Piatetsky-Shapiro and P. Tamayo, "Microarray data mining: facing the challenges," *ACM SIGKDD Explorations Newsletter*, vol. 5, no. 2, pp. 1–5, 2003.

[10] A. Lommen, "Metalign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing," *Analytical chemistry*, vol. 81, no. 8, pp. 3079–3086, 2009.

[11] A. Holzinger, M. Dehmer, and I. Jurisica, "Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions," *BMC bioinformatics*, vol. 15, no. 6, p. I1, 2014.

[12] M. Wilkins, "Proteomics data mining," *Expert review of proteomics*, vol. 6, no. 6, pp. 599–603, 2009.

[13] D. Teodoro, R. Choquet, E. Pasche, J. Gobeill, C. Daniel, P. Ruch, and C. Lovis, "Biomedical data management: a proposal framework.," in *MIE*, pp. 175–179, Citeseer, 2009.

[14] M. Y. Galperin, "The molecular biology database collection: 2008 update," *Nucleic Acids Research*, vol. 36, no. suppl1, pp. D2–D4, 2008.

[15] A. Sturn, J. Quackenbush, and Z. Trajanoski, "Genesis: cluster analysis of microarray data," *Bioinformatics*, vol. 18, no. 1, pp. 207–208, 2002.

[16] A. Karnovsky, T. Weymouth, T. Hull, V. G. Tarcea, G. Scardoni, C. Laudanna, M. A. Sartor, K. A. Stringer, H. Jagadish, C. Burant, *et al.*, "Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data," *Bioinformatics*, vol. 28, no. 3, pp. 373–380, 2011.

[17] D. Tabas-Madrid, R. Nogales-Cadenas, and A. Pascual-Montano, "Genecodis3: a non-redundant and modular enrichment analysis tool for functional genomics," *Nucleic acids research*, vol. 40, no. W1, pp. W478–W483, 2012.

[18] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, "G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences," *Behavior research methods*, vol. 39, no. 2, pp. 175–191, 2007.

[19] R. P. Nair, K. C. Duffin, C. Helms, J. Ding, P. E. Stuart, D. Goldgar, J. E. Gudjonsson, Y. Li, T. Tejasvi, B.-J. Feng, *et al.*, "Genome-wide scan reveals association of psoriasis with il-23 and nf-κb pathways," *Nature genetics*, vol. 41, no. 2, pp. 199–204, 2009.

[20] M. Suárez-Farinas, K. Li, J. Fuentes-Duculan, K. Hayden, C. Brodmerkel, and J. G. Krueger, "Expanding the psoriasis disease profile: interrogation of the skin and serum of patients with moderate-to-severe psoriasis," *Journal of Investigative Dermatology*, vol. 132, no. 11, pp. 2552–2564, 2012.

[21] J. Bigler, H. A. Rand, K. Kerkof, M. Timour, and C. B. Russell, "Cross-study homogeneity of psoriasis gene expression in skin across a large expression range," *PLoS One*, vol. 8, no. 1, p. e52242, 2013.

[22] J. Kim, R. Bissonnette, J. Lee, J. C. da Rosa, M. Suárez-Fariñas, M. A. Lowes, and J. G. Krueger, "The spectrum of mild to severe psoriasis vulgaris is defined by a common activation of il-17 pathway genes, but with key differences in immune regulatory genes," *Journal of Investigative Dermatology*, vol. 136, no. 11, pp. 2173–2182, 2016.

[23] Y. Yao, L. Richman, C. Morehouse, M. De Los Reyes, B. W. Higgs, A. Boutrin, B. White, A. Coyle, J. Krueger, P. A. Kiener, *et al.*, "Type i interferon: potential therapeutic target for psoriasis?," *PloS one*, vol. 3, no. 7, p. e2737, 2008.

[24] M. Suárez-Fariñas, S. J. Tintle, A. Shemer, A. Chiricozzi, K. Nograles, I. Cardinale, S. Duan, A. M. Bowcock, J. G. Krueger, and E. Guttman-Yassky, "Nonlesional atopic dermatitis skin is characterized by broad terminal differentiation defects and variable immune abnormalities," *Journal of Allergy and Clinical Immunology*, vol. 127, no. 4, pp. 954–964, 2011.

[25] S. Tintle, A. Shemer, M. Suárez-Fariñas, H. Fujita, P. Gilleaudeau, M. Sullivan-Whalen, L. Johnson-Huang, A. Chiricozzi, I. Cardinale, S. Duan, *et al.*, "Reversal of atopic dermatitis with narrow-band uvb phototherapy and biomarkers for therapeutic response," *Journal of Allergy and Clinical Immunology*, vol. 128, no. 3, pp. 583–593, 2011.

[26] J. K. Gittler, A. Shemer, M. Suárez-Fariñas, J. Fuentes-Duculan, K. J. Gulewicz, C. Q. Wang, H. Mitsui, I. Cardinale, C. de Guzman Strong, J. G. Krueger, *et al.*, "Progressive activation of t h 2/t h 22 cytokines and selective epidermal proteins characterizes acute and chronic atopic dermatitis," *Journal of Allergy and Clinical Immunology*, vol. 130, no. 6, pp. 1344–1354, 2012.

[27] S. R. Rapp, S. R. Feldman, M. L. Exum, A. B. Fleischer, and D. M. Reboussin, "Psoriasis causes as much disability as other major medical diseases," *Journal of the American Academy of Dermatology*, vol. 41, no. 3, pp. 401–407, 1999.

[28] S. Jowett and T. Ryan, "Skin disease and handicap: an analysis of the impact of skin conditions," *Social science & medicine*, vol. 20, no. 4, pp. 425–429, 1985.

[29] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: Ncbi gene expression and hybridization array data repository," *Nucleic acids research*, vol. 30, no. 1, pp. 207–210, 2002.

[30] A. L. Dixon, L. Liang, M. F. Moffatt, W. Chen, S. Heath, K. C. Wong, J. Taylor, E. Burnett, I. Gut, M. Farrall, *et al.*, "A genome-wide association study of global gene expression," *Nature genetics*, vol. 39, no. 10, p. 1202, 2007.

[31] E. D. Roberson and A. M. Bowcock, "Psoriasis genetics: breaking the barrier," *Trends in Genetics*, vol. 26, no. 9, pp. 415–423, 2010.

[32] I. Leigh, H. Navsaria, P. Purkis, I. McKay, P. Bowden, and P. Riddle, "Keratins (kl6 and kl7) as markers of keratinocyte hyperproliferation in psoriasis in vivo and in vitro," *British Journal of Dermatology*, vol. 133, no. 4, pp. 501–511, 1995.

[33] J. Vegfors, S. Petersson, A. Kovacs, K. Polyak, and C. Enerbäck, "The expression of psoriasin (s100a7) and cd24 is linked and related to the differentiation of mammary epithelial cells," *PloS one*, vol. 7, no. 12, p. e53119, 2012.

[34] K. Kainu, K. Kivinen, M. Zucchelli, S. Suomela, J. Kere, A. Inerot, B. S. Baker, A. V. Powles, L. Fry, L. Samuelsson, *et al.*, "Association of psoriasis to pglyrp and sprr genes at psors4 locus on 1q shows heterogeneity between finnish, swedish and irish families," *Experimental dermatology*, vol. 18, no. 2, pp. 109–115, 2009.

[35] J. G. Bergboer, P. L. Zeeuwen, and J. Schalkwijk, "Genetics of psoriasis: evidence for epistatic interaction between skin barrier abnormalities and immune deviation," *Journal of Investigative Dermatology*, vol. 132, no. 10, pp. 2320–2331, 2012.

[36] H. Michibata, H. Chiba, K. Wakimoto, M. Seishima, S. Kawasaki, K. Okubo, H. Mitsui, H. Torii, and Y. Imai, "Identification and characterization of a novel component of the cornified envelope, cornifelin," *Biochemical and biophysical research communications*, vol. 318, no. 4, pp. 803–813, 2004.

[37] S. Marrakchi, P. Guigue, B. R. Renshaw, A. Puel, X.-Y. Pei, S. Fraitag, J. Zribi, E. Bal, C. Cluzeau, M. Chrabieh, *et al.*, "Interleukin-36–receptor antagonist deficiency and generalized pustular psoriasis," *New England Journal of Medicine*, vol. 365, no. 7, pp. 620–628, 2011.

[38] H. Niehues, L. C. Tsoi, D. A. van der Krieken, P. A. Jansen, M. A. Oortveld, D. Rodijk-Olthuis, I. M. van Vlijmen, W. J. Hendriks, R. W. Helder, J. A. Bouwstra, *et al.*, "Psoriasis-associated late cornified envelope (lce) proteins have antibacterial activity," *Journal of Investigative Dermatology*, vol. 137, no. 11, pp. 2380–2388, 2017.

[39] S. Jiang, T. E. Hinchliffe, and T. Wu, "Biomarkers of an autoimmune skin disease—psoriasis," *Genomics, proteomics & bioinformatics*, vol. 13, no. 4, pp. 224–233, 2015.

[40] L.-D. Sun, H. Cheng, Z.-X. Wang, A.-P. Zhang, P.-G. Wang, J.-H. Xu, Q.-X. Zhu, H.-S. Zhou, E. Ellinghaus, F.-R. Zhang, *et al.*, "Association analyses identify six new psoriasis susceptibility loci in the chinese population," *Nature genetics*, vol. 42, no. 11, p. 1005, 2010.

[41] A. R. Djalilian, D. McGaughey, S. Patel, E. Y. Seo, C. Yang, J. Cheng, M. Tomic, S. Sinha, A. Ishida-Yamamoto, and J. A. Segre, "Connexin 26 regulates epidermal barrier and wound remodeling and promotes psoriasiform response," *The Journal of clinical investigation*, vol. 116, no. 5, pp. 1243–1253, 2006.

[42] B. Algermissen, J. Sitzmann, P. LeMotte, and B. Czarnetzki, "Differential expression of crabp ii, psoriasin and cytokeratin 1 mrna in human skin diseases," *Archives of dermatological research*, vol. 288, no. 8, p. 426, 1996.

[43] D. Bruch-Gerharz, O. Schnorr, C. Suschek, K.-F. Beck, J. Pfeilschifter, T. Ruzicka, and V. Kolb-Bachofen, "Arginase 1 overexpression in psoriasis: limitation of inducible nitric oxide synthase activity as a molecular mechanism for keratinocyte hyperproliferation," *The American journal of pathology*, vol. 162, no. 1, pp. 203–211, 2003.

[44] F. O. Nestle, C. Conrad, A. Tun-Kyi, B. Homey, M. Gombert, O. Boyman, G. Burg, Y.-J. Liu, and M. Gilliet, "Plasmacytoid predendritic cells initiate psoriasis through interferon-α production," *Journal of Experimental Medicine*, vol. 202, no. 1, pp. 135–143, 2005.

[45] S. Suomela, L. Cao, A. Bowcock, and U. Saarialho-Kere, "Interferon α-inducible protein 27 (ifi27) is upregulated in psoriatic skin and certain epithelial cancers," *Journal of Investigative Dermatology*, vol. 122, no. 3, pp. 717–721, 2004.

[46] A. Chiricozzi, E. Guttman-Yassky, M. Suárez-Farinas, K. E. Nograles, S. Tian, I. Cardinale, S. Chimenti, and J. G. Krueger, "Integrative responses to il-17 and tnf-α in human keratinocytes account for key inflammatory pathogenic circuits in psoriasis," *Journal of Investigative Dermatology*, vol. 131, no. 3, pp. 677–687, 2011.

[47] M. A. Lowes, T. Kikuchi, J. Fuentes-Duculan, I. Cardinale, L. C. Zaba, A. S. Haider, E. P. Bowman, and J. G. Krueger, "Psoriasis vulgaris lesions contain discrete populations of th1 and th17 t cells," *Journal of Investigative Dermatology*, vol. 128, no. 5, pp. 1207–1211, 2008.

[48] P. A. Jansen, D. Rodijk-Olthuis, E. J. Hollox, M. Kamsteeg, G. S. Tjabringa, G. J. de Jongh, I. M. van Vlijmen-Willems, J. G. Bergboer, M. M. van Rossum, E. M. de Jong, *et al.*, "β-defensin-2 protein is a serum biomarker for disease activity in psoriasis and reaches biologically relevant concentrations in lesional skin," *PloS one*, vol. 4, no. 3, p. e4725, 2009.

[49] J. E. Gudjonsson, J. Ding, A. Johnston, T. Tejasvi, A. M. Guzman, R. P. Nair, J. J. Voorhees, G. R. Abecasis, and J. T. Elder, "Assessment of the psoriatic transcriptome in a large sample: additional regulated genes and comparisons with in vitro models," *Journal of Investigative Dermatology*, vol. 130, no. 7, pp. 1829–1840, 2010.

[50] L. M. Roesner, P. Kienlin, G. Begemann, O. Dittrich-Breiholz, and T. Werfel, "Inflammatory marker analysis in psoriatic skin under topical phosphodiesterase 4 inhibitor treatment," *Journal of Allergy and Clinical Immunology*, vol. 140, no. 4, pp. 1184–1187, 2017.

A Biomedical relation graphs

In this appendix the graphs for biomedical relations are shown. These graphs did not show any significant results, however do show how the method worked.

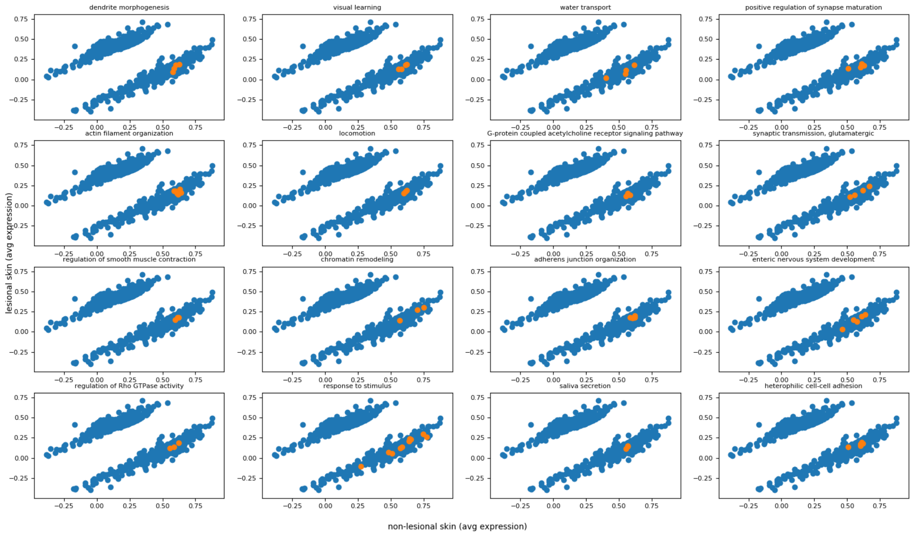


Figure 8: The 16 processes that showed the highest difference in gene expression between lesional and non-lesional data. The data is standardized by feature and orange points are genes that are related to the process, Blue dots are genes unrelated to the process

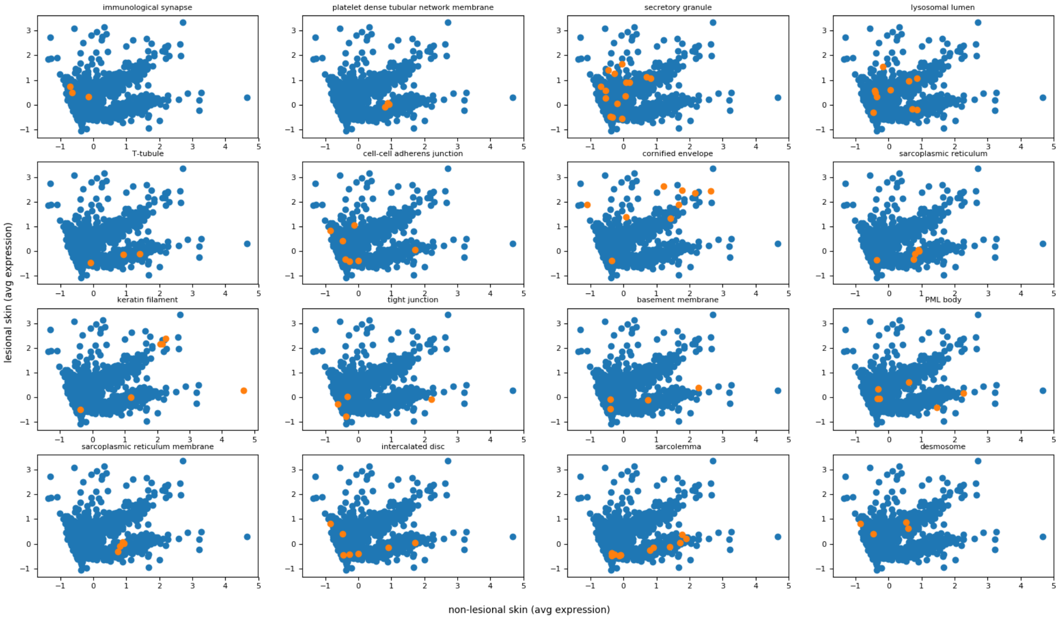


Figure 9: The 16 cellular locations that showed the highest difference in gene expression between lesional and non-lesional data. The data is standardized by sample and orange points are genes that are related to the cellular location, Blue dots are genes unrelated to the cellular location

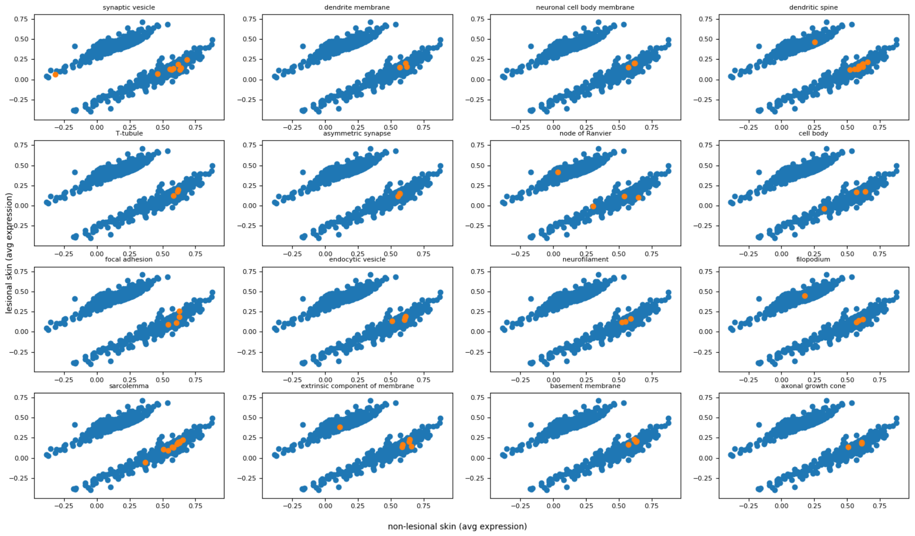


Figure 10: The 16 cellular locations that showed the highest difference in gene expression between lesional and non-lesional data. The data is standardized by feature and orange points are genes that are related to the cellular location, Blue dots are genes unrelated to the cellular location

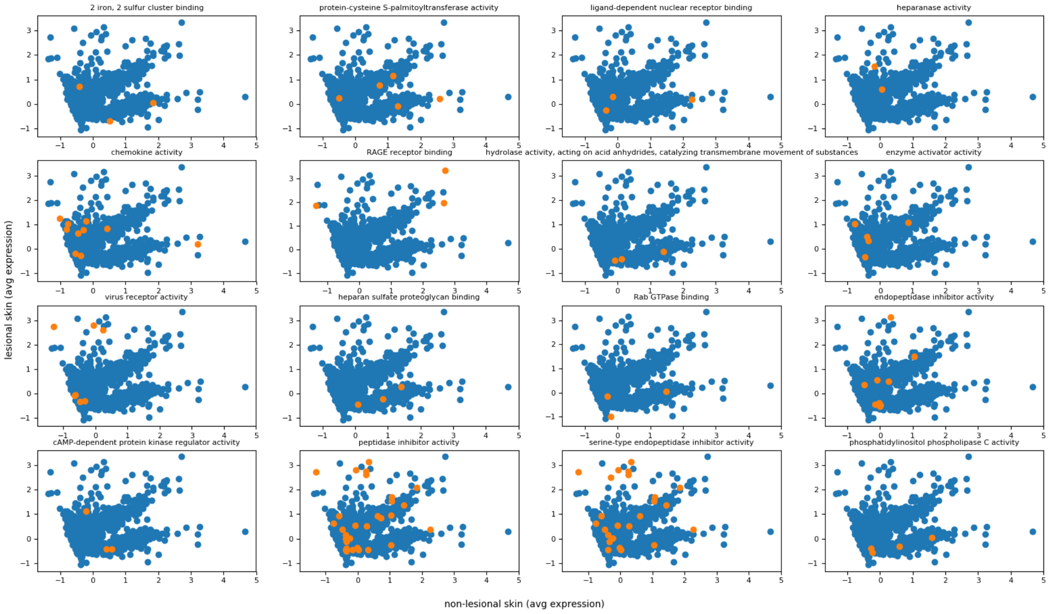


Figure 11: The 16 molecular relations that showed the highest difference in gene expression between lesional and non-lesional data. The data is standardized by sample and orange points are genes that are related to the molecular relation, Blue dots are genes unrelated to the molecular relation

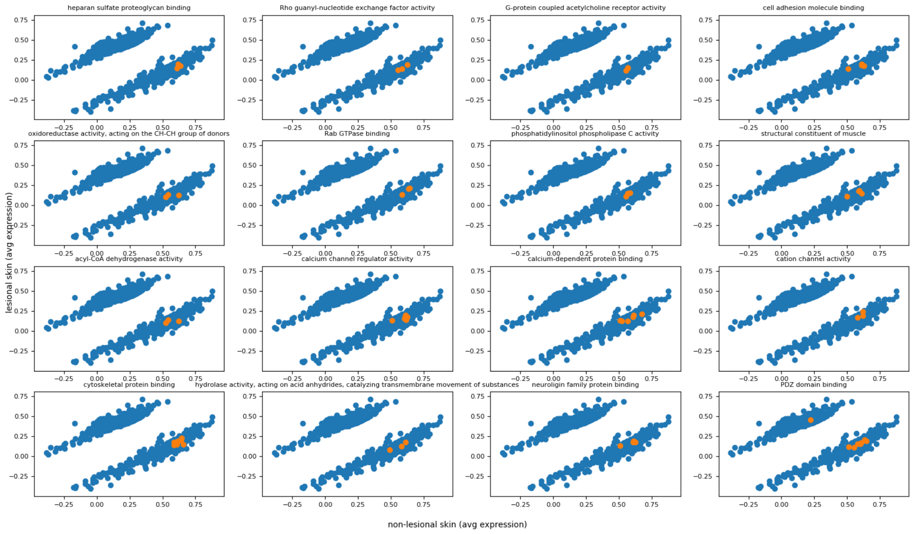


Figure 12: The 16 molecular reactions that showed the highest difference in gene expression between lesional and non-lesional data. The data is standardized by feature and orange points are genes that are related to the molecular relation, Blue dots are genes unrelated to the molecular relation

B Cluster Gene Specifications

In this appendix the genes from cluster 5 (Table 8) and cluster 8 (Table 9) are listed. Their commonly known name and a their possible link to psoriasis both are explained, if found in literature.

Table 8: All genes in the highlighted cluster 5 with their link to psoriasis

Gene name	Gene description	Role in psoriasis
NM_006945	small proline-rich protein 2B	skin barrier development[31]
AF061812	keratin 16	markers of keratinocyte hyperproliferation[32]
BG327863	CD24 molecule	differentiation marker of epithelial cells[33]
BF575466	small proline-rich protein 3	skin barrier development[34]
AI923984	small proline-rich protein 1A	skin barrier development[35]
AB049591	cornifelin	Related to cornified envelope[36]
AF216693	interleukin 36 receptor antagonist	Antinflammatory cytokine//Aberrant interleukin-36Ra structure and function lead to unregulated secretion of inflammatory cytokines and generalized pustular psoriasis[37]
AB048288	late cornified envelope 3D	Late cornified envelope (LCE) genes, located in the epidermal differentiation complex on chromosome 1, encode a family of 18 proteins of unknown function, whose expression is largely restricted to epidermis[38]
AW238654	S100 calcium binding protein A8	proteins that attract leukocytes, upregulated in psoriatic keratinocytes, also synthesized by neutrophils or myeloid dendritic cells that are recruited into psoriatic epidermis[31]
J00269	keratin 6A	hyperproliferation marker[39]
NM_002964	S100 calcium binding protein A8	proteins that attract leukocytes, upregulated in psoriatic keratinocytes, also synthesized by neutrophils or myeloid dendritic cells that are recruited into psoriatic epidermis[31]
NM_005987	small proline-rich protein 1A	skin barrier development[35]
L42612	keratin 6B	hyperproliferation marker[39]
AL569511	keratin 6A /// keratin 6B /// keratin 6C	hyperproliferation marker[39]
AJ001698	serpin peptidase inhibitor clade B (ovalbumin) member 13	The encoded protein inhibits the activity of cathepsin K and is itself transcriptionally repressed by RUNX1. Known to be upregulated in lesional skin.[20]
AI286239	No name	No connection found with psoriasis
M86849	gap junction protein beta 2 26kDa	Associated with psoriasis, regulates epidermal barrier and wound remodeling[40, 41]
NM_001878	cellular retinoic acid binding protein 2	Altered epidermal gene expression of CRABP II is not disease-specific and may reflect instead an altered state of epidermal differentiation and/or may be linked to the inflammation and cellular infiltration common various conditions.[42]
NM_000045	arginase 1	Participates in the regulation of iNOS activity by competing for the common substrate l-arginine, is highly overexpressed in the hyperproliferative psoriatic epidermis and is co-expressed with iNOS[43]
NM_005532	interferon, alpha-inducible protein 27	drives the development of psoriasis, marker of epithelial proliferation[44, 45]
NM_003125	small proline-rich protein 1B	skin barrier development[35]
NM_005130	fibroblast growth factor binding protein 1	FGFBP2 is downregulated, FGFBP1 not mentioned[20]
NM_002108	histidine ammonia-lyase	No connection found with psoriasis

Table 9: All genes in the highlighted cluster 8 with their link to psoriasis

Gene name	Gene description	Role in psoriasis
AJ243672	S100 calcium binding protein A7A	IL-17 regulated inflammatory gene[46]
U19557	serpin peptidase inhibitor clade B (ovalbumin) member 3 /// serpin peptidase inhibitor, clade B (ovalbumin), member 4	Immune related gene and biomarker of psoriasis[47, 39]
NM_002965	S100 calcium binding protein A9	innate immune mediators, IL-17 regulated gene[46]
BC005224	serpin peptidase inhibitor clade B (ovalbumin) member 3	Immune related gene and biomarker of psoriasis[47, 39]
U19556	serpin peptidase inhibitor clade B (ovalbumin) member 3	Immune related gene and biomarker of psoriasis[47, 39]
NM_004942	defensin beta 4A /// defensin beta 4B	antibiotic peptide which is locally regulated by inflammation[48]
L10343	peptidase inhibitor 3 skin-derived	Upregulated in psoriasis, his gene encodes an elastase-specific inhibitor that functions as an antimicrobial peptide against Gram-positive and Gram-negative bacteria, and fungal pathogens.[49, 50]
NM_002638	peptidase inhibitor 3 skin-derived	Upregulated in psoriasis, his gene encodes an elastase-specific inhibitor that functions as an antimicrobial peptide against Gram-positive and Gram-negative bacteria, and fungal pathogens.[49, 50]