12 Knowledge Discovery with DBS

In contrast to chapter 11, in which Databionic swarm (DBS) clustering was applied to recognize more or less obvious knowledge, this chapter shows that DBS is also able to discover new knowledge. A hydrological data set of multivariate time series [Aubert et al., 2016] and a data set consisting of pain genes [Ultsch et al., 2016b] are used for this purpose. In [Aubert et al., 2016], a high-frequency time series analysis was performed, but no prediction could be made. Here, the focus is placed on daily frequency.
The analysis of [Ultsch et al., 2016b] concentrated on chronic pain, and for that reason, it required searching for candidate genes that modulate pain chronification. This chapter, however, focuses on defining the distances between genes and grouping genes by semantic similarity, which can be explained based on overrepresentation analysis (ORA) [Backes et al., 2007].

12.1 Hydrology

> *"Human activities modify the global nitrogen cycle, particularly through farming. These practices have unintended consequences; for example, nitrate lost from terrestrial runoff to streams and estuaries can impact aquatic life" [Aubert et al., 2016].*

A greater understanding of water quality variations can improve the evaluation of the state of water bodies and lead to better recommendations for appropriate and efficient management practices [Cirmo/McDonnell, 1997]**.** Accordingly, the objective here is to predict water quality in the Schwingbach catchment[73] using the currently available variables related to chemical water quality: nitrate and (electrical) conductivity (*N&C*) which is a part of the science of hydrology. Electrical conductivity is a measure that reflects the water quality as a whole; this is because it indicates the variations in the presence of ions other than nitrate in the water body [Aubert, 2015]. Nitrate in water bodies is partially responsible for the phenomenon of eutrophication [Diaz, 2001]. Eutrophication occurs when an excess of nutrients (i.e., nitrate) leads to uncontrollable growth of aquatic plant life, followed by a depletion of the dissolved oxygen [Diaz, 2001; Howarth et al., 1996]. For this reason, the nitrate concentration is one of the parameters used to evaluate water quality.

> *"The available dataset contained in total 32,196 data points for each of the 14 variables (in total, 4% missing data). For technical reasons, no nitrate data were available during winter, so the actual time span of nitrate monitoring was 05 March 2013 12:45 to 24 September 2013 12:30 and 27 April 2014 00:00 to 23 October 13:15. Data were analyzed as a whole, without differentiating between the hydrological years"* [Aubert et al., 2016].

Conductivity, in particular, will be explained using another set of variables, which are indicators of hydrological and biological conditions. In contrast to the temporal high-frequency analysis (with 15-minute intervals) of [Aubert et al., 2016], here, the daily courses for each variable were calculated as the sums of all daily measurements, resulting in a low-frequency analysis. The missing values were imputed using the seven-nearest-neighbors approach. All variables were linearly decorrelated, and the logarithms of the variables q13 and q18 were calculated. Subsequently, all variables, with the exception of rain, were normalized to values between zero

[73] A catchment is a dynamic system, and current observations depend on previous hydrological states [Aubert et al., 2016].

M. C. Thrun, *Projection-Based Clustering through Self-Organization and Swarm Intelligence*, https://doi.org/10.1007/978-3-658-20540-9_12

and one through robust normalization. The outliers in the rain variable were detected via ABC analysis [Ultsch/Lötsch, 2015]: in the ABC analysis, rain was normalized with respect to the minimum value in group A and then all points in group A were set to a value of 1.1 for rain, and. After feature selection the data set had in 12 variables over 343 days.

The preprocessed daily courses are shown in Figure 12.1. The preprocessing resulted in Euclidean distances with a multimodal distribution (Figure 12.2). The first mode represents the intracluster distances, and the second mode represents the intercluster distances (see also chapter 3, Figure 3.1).

DBS was used for visualization and clustering. The outliers were marked interactively, resulting in five classes (Figure 12.4). The clusters have small intracluster distances and high intercluster distances, as visualized using DBS (Figure 12.4) and confirmed by the heatmap (Figure 12.4). The silhouette plot shows that all clusters can be well modeled as hyperspheres (Figure 12.3).

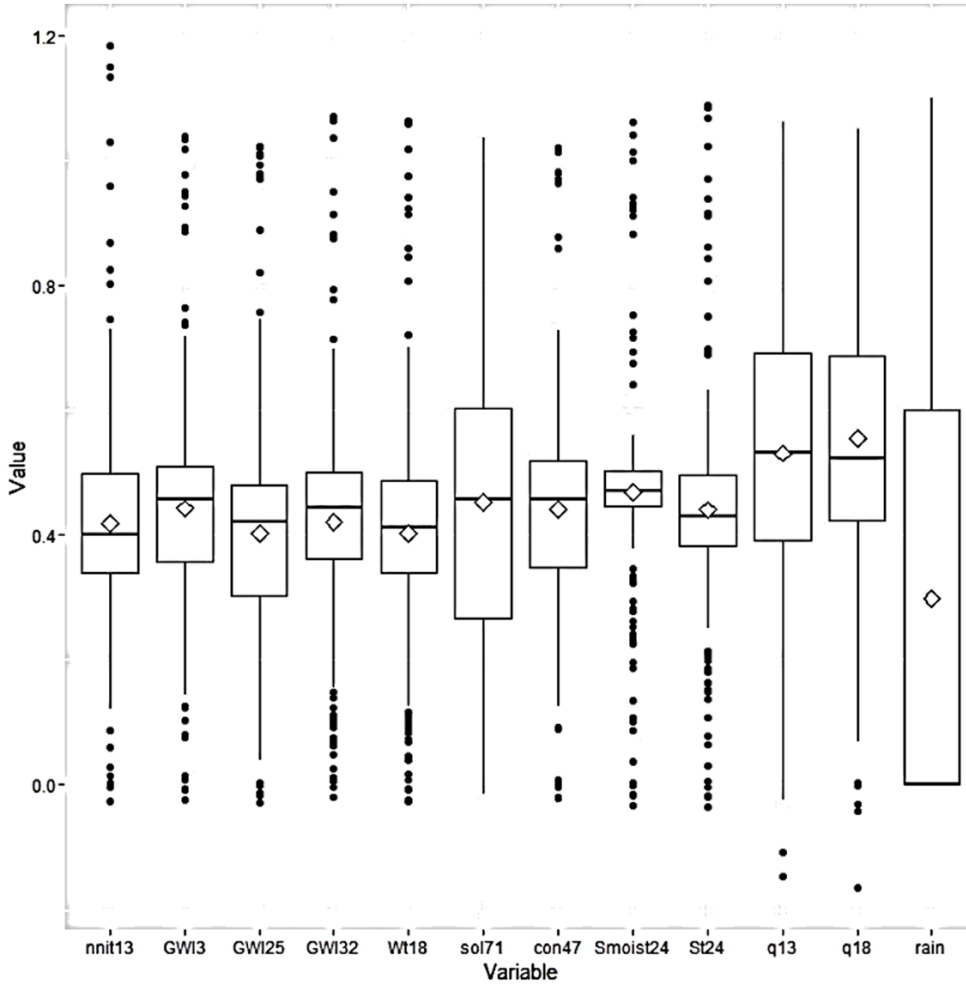

Figure 12.1: Variances of variables after preprocessing and feature extraction visualized using boxplots after the preprocessing of the hydrology data set.

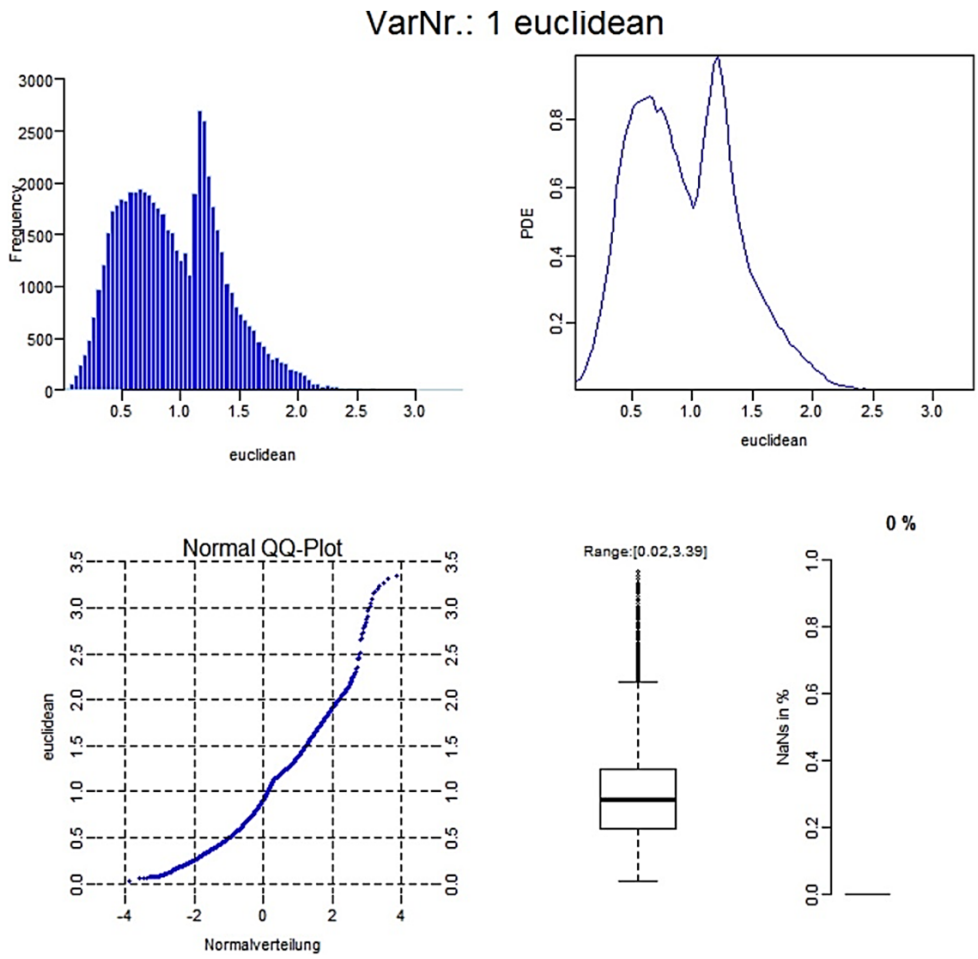

Figure 12.2: Distribution analysis of the distances. The first mode represents the intracluster distances, and the second mode represents the intercluster distances (for further explanation see chapter 3, Figure 3.1).

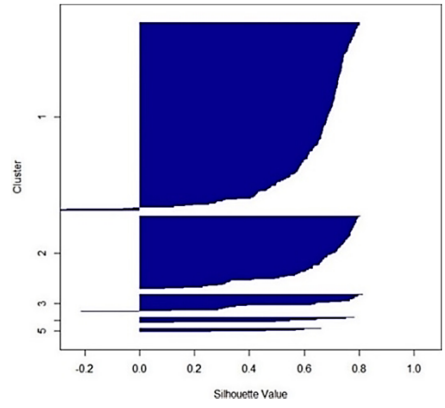

Figure 12.3: Silhouette plot of the DBS clustering set indicates that data points (y-axis) above a value of 0.5 (x-axis) have been assigned to an appropriate cluster.

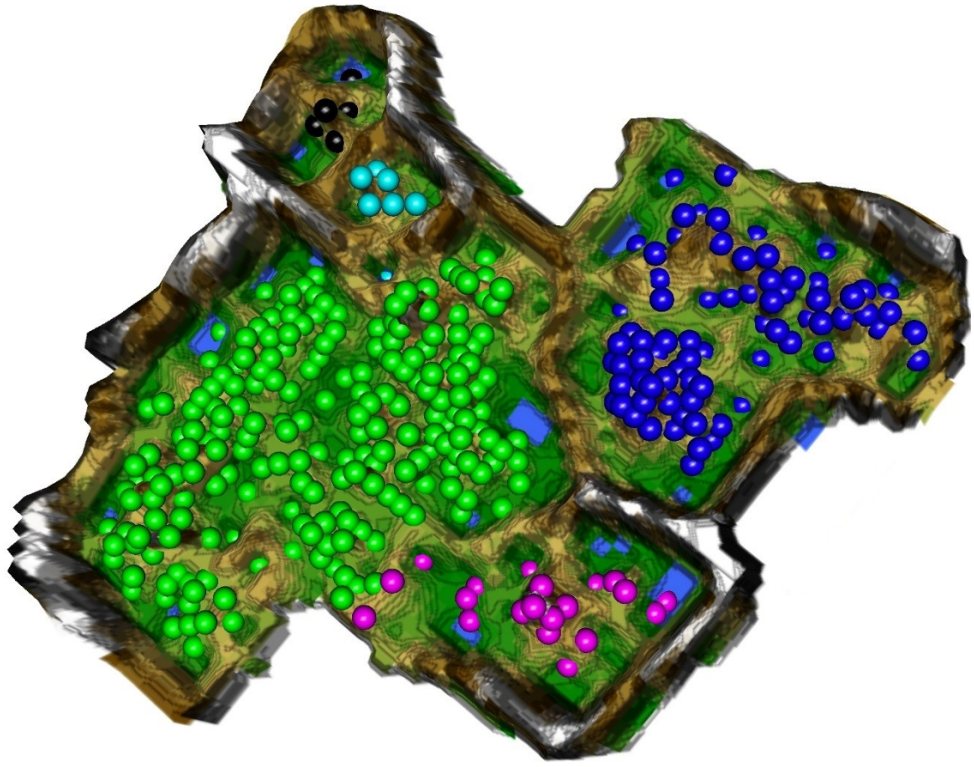

Figure 12.4: Five clusters are shown in the topographic map of DBS of the Hydrology data set. For 3D print see supplement G, Figure G.24.

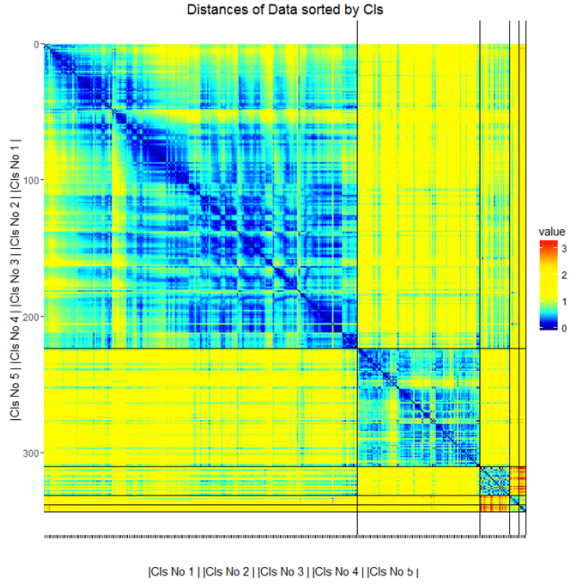

Figure 12.5: The five clusters have clearly distinctive distances, as shown by the heatmap; there are small distances within each cluster and large distances between the clusters.

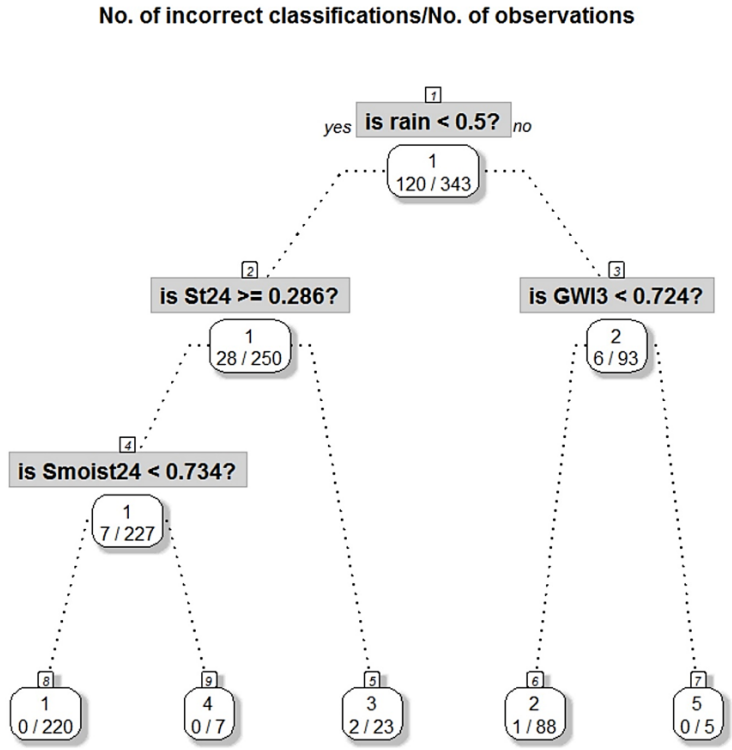

Figure 12.6: Classification and Regression Tree (CART) analysis rules for the hydrology data set with the five clusters identified by DBS. Applying the rules to the clustering combined with the data set results in three misclassified points (0.9%). Abbreviations: rainfall intensity (rain), soil temperature (St24), soil moisture (Smoist24), groundwater level at point 3 (GWl3). All values are expressed as percentages.

12.1.1 Knowledge Acquisition and Prediction in the Hydrology Data Set

Here, the rules extracted from the Classification and Regression Tree (CART) decision tree, as shown in Figure 12.6, were applied to the clustering. In comparison to the DBS clustering, the application of the CART rules to the data set results in the misclassification of three data points (0.9%). Based on this finding, it can be said that the rules precisely classify the data set (Figure 12.6). The generated rules are listed in Table 12.1.

Table 12.1: The CART rules based on Figure 12.6, in which the clusters of Figure 12.4 are used. Abbreviations: rainfall intensity (rain), soil temperature (St24), soil moisture (Smoist24), groundwater level at point 3 (GWl3). All values are expressed as percentages.

Rule No.	DBS Cluster No.	No. of Days	Rule
R1	1	223	if rain < 0.5 and St24 ≥ 0.29 and Smoist24 < 0.73
R2	4	7	if rain < 0.5 and St24≥ 0.29 and Smoist24 ≥ 0.73
R3	3	21	if rain < 0.5 and St24 < 0.29
R4	2	87	if rain ≥ 0.5 and GWl3 < 0.72
R5	5	5	if rain ≥ 0.5 and GWl3 ≥ 0.72

The N&C measurements can be described by two variables related to biological processes, namely, soil temperature and soil moisture, and two variables related to hydrological processes, namely, rainfall intensity and groundwater level at point 3, which represents downslope conditions. Temperature influences the activities of living organisms, such as soil microbial organisms [Zak et al., 1999]. Soil moisture determines microbial activities, such as long-term inactivity in dried soil followed by wetting [Borken/Matzner, 2009]. The groundwater level (or head, in m) is the main factor driving discharge in a catchment [Orlowski et al., 2014]. Rainfall intensity triggers discharge and affects soil moisture as well as leaching of nutrients [Orlowski et al., 2014].

A thorough examination of the CART results based on the five distinguishing rules R (Tab. 1) yields the following classes C:

C1/R1: Low rain, higher soil temperature, lower soil moisture => *DryDays WetHotGround*
C2/R4: High rain, lower downslope groundwater level => *Rain Shower*
C3/R3: Low rain, low soil temperature => *DryDays Cold Ground*
C4/R2: Low rain, higher soil temperature, high soil moisture => *DryDays DryHotGround*
C5/R5: High rain, high downslope groundwater level => *Rainy Days*

With regard to N&C, these classes can be distinguished as follows: the first two classes (green and blue) are responsible for normal N&C, the third class (magenta) is associated with low N&C, and the fourth and fifth classes (teal and black) are responsible for high N&C (Figure 12.7).

After a rain shower or on dry days when the ground is wet and hot, the N&C concentrations are normal. The N&C concentrations are high (above 50%) on rainy days, when the downslope groundwater level is above 72%. The N&C concentration is low (<25%) on dry days (below 50% rain) when the ground is cold (below 29% of the maximum ground temperature). These definitions enable future predictions of daily N&C concentrations.

It is assumed here that the structures associated with the 5 clusters described by these classes are defined by discontinuities. Consequently, the clusters should contain samples of different natures and based on different processes. Given this assumption, it is valid to statistically test whether the N&C distributions significantly differ between clusters. The Kolmogorov–Smirnov test (KS test) is a nonparametric two-sample test of the null hypothesis that two variables are drawn from the same continuous distribution [Conover, 1971, pp. 309-314], and it is implemented in the R language [R Development Core Team, 2008].

The statistical results are shown in supplement F, Tab. 1 and 2. All N&C distributions significantly differ between clusters, with the exception of cluster 4 compared with 5, for both variables.

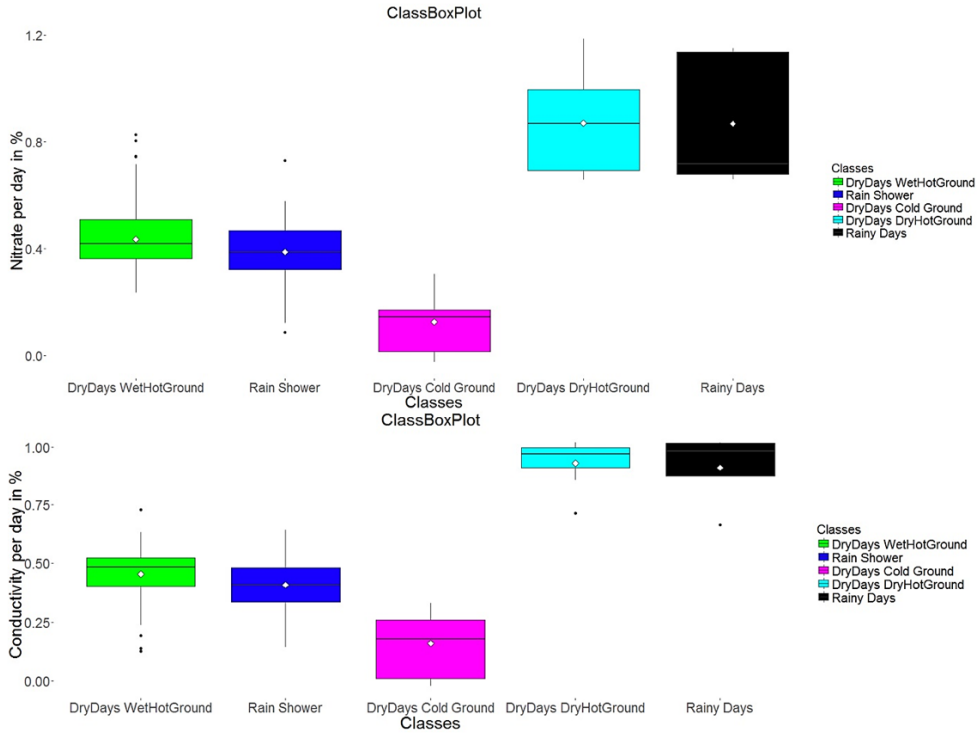

Figure 12.7: Boxplots of the five classes with regard to nitrate N (top) and conductivity C (bottom). All values are expressed as percentages.

12.2 Pain Genes

In [Ultsch et al., 2016b], a set of genes with relevance to pain[74] was obtained from four sources, and the search of several databases and studies (e.g., the Pain Genes Database, the PubMed database) was described in detail. This search yielded a set of n = 535 genes, subsequently referred to as *pain genes* in [Ultsch et al., 2016b].

After accessing the Gene Ontology (GO) database in this work, 528 of the pain genes were found to be annotated, and the remaining seven genes were disregarded in the subsequent analysis (feature selection). Various types of annotation (evidence codes) are possible. When the inverse document frequency *idf* is used [Sparck Jones, 1972], the distances between these genes are defined as follows (as discussed in [Ultsch, 2014b]):

Let the documents be represented by GO terms T, and let the terms used to calculate *idf* be represented by the genes G, which are coded with numbers defined by the National Center for Biotechnology Information (NCBI) [NCBI, 2013]; the term frequency *tf* is then the frequency of occurrence of a gene in a given document divided by the maximal occurrence of the gene in any document:

[74] "An unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage" [Merskey/Bogduk, 1994].

$$tf(G,T) = \frac{f(G,T)}{\max(f)} \qquad (1)$$

If only manually curated evidence codes are used for annotation, then $tf(G,T) = 1$.
Let N be the number of GO terms to which the pain genes are annotated, and let n_i be the number of GO terms to which a pain gene with a given NCBI number is annotated; then, the inverse document frequency is defined as

$$idf_i = log\left(1 + \frac{N}{n_i}\right) \qquad (2)$$

and the term frequency–inverse document frequency is defined as

$$tfidf = tf(G,T) * idf_i = 1 * idf_i \qquad (3)$$

A gene that is annotated to only some GO terms is more meaningful than one that is annotated to almost every or only a few GO terms . Hence, the inverse document frequency reduces the weights of genes that occur very frequently among the GO terms and increases the weight of genes that occur rarely. The distance D between two genes l and j is defined as the absolute distance in terms of *idf*:

$$D(l,j) = abs(idf_l - idf_j) \qquad (4)$$

This distance was used to generate the DBS visualization shown in Figure 12.9, and clustering was automatically performed after the identification of 8 clusters in the visualization. The clusters are verified by the heatmap presented in Figure 12.10 and the Silhouette plot in Figure 12.8.

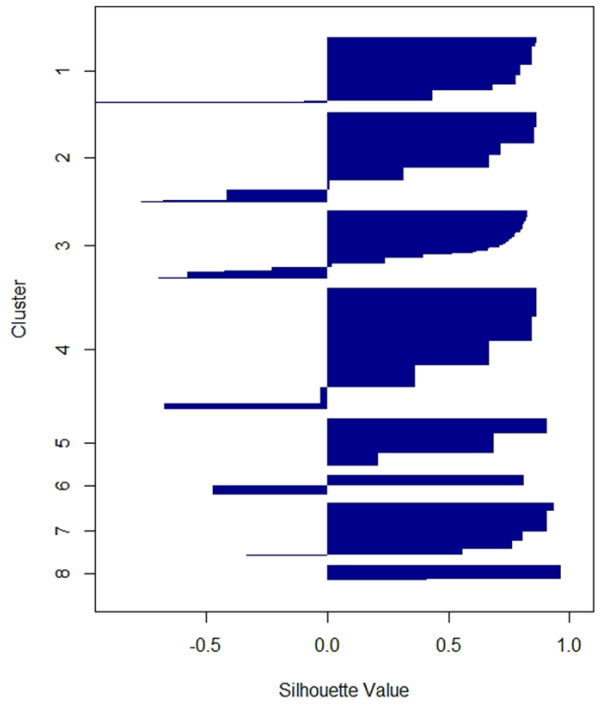

Figure 12.8: Silhouette plot of the DBS clustering of pain genes. Most of clusters of pain genes can be modeled as hyperspheres. However, cluster 6 has a different high-dimensional structure.

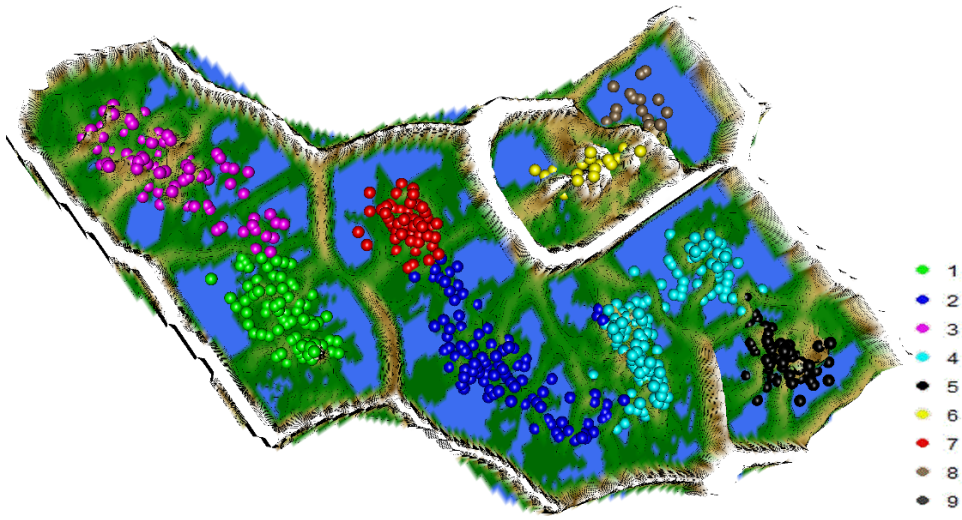

Figure 12.9: Topographic map of DBS clustering of 528 pain genes. Clusters 1 and 3 and clusters 2 and 4 are very similar to each other. Cluster 6, labeled in yellow, consists of outliers. The counts per cluster, from 1 to 8, are 72, 99, 75, 133, 53, 21, 58, and 17. For 3D print see supplement G, Figure G.25.

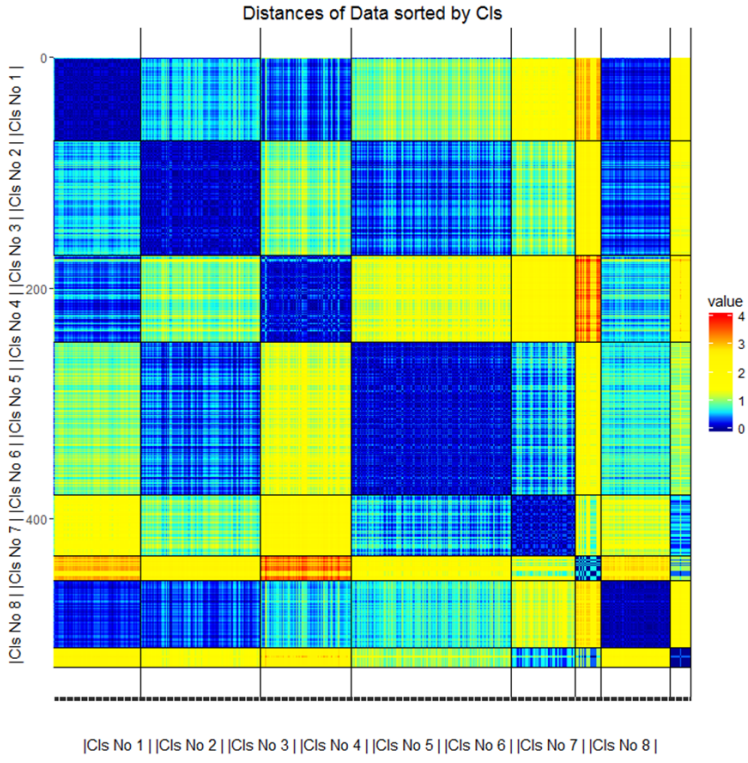

Figure 12.10: Heatmap of the distances with regard to the 8 identified clusters of pain genes, which verifies that the clustering is sound. Clusters 1 and 3 and clusters 2 and 4 are very similar to each other. Cluster 6 is clearly defined by outliers.

12.2.1 Prior Knowledge

The pain genes were analyzed by means of ORA, revealing several important functions, as listed below. If the distance definition and DBS clustering were applied correctly to the pain genes data set, it should be possible to rediscover structures that are already known from two main publications on this topic. [Lötsch et al., 2013] defined twelve functions of pain for 460 pain genes (Figure 12.11) [Lötsch et al., 2013]:

1.) regulation of localization
2.) behavior
3.) response to wounding
4.) response to organic substance
5.) cellular ion homeostasis
6.) ion transport
7.) synaptic transmission
8.) G protein-coupled receptor protein signaling pathway
9.) intracellular signal transduction
10.) positive regulation of biological process
11.) regulation of system process
12.) anatomical structure development

Additionally, in 2016, twelve chronification functions of 535 pain genes were identified [Ultsch et al., 2016b]:

1.) single-organism cellular process
2.) biological regulation
3.) cell communication
4.) cellular response to stimulus
5.) localization
6.) response to stress
7.) phosphorus metabolic process
8.) nervous system development
9.) cell death
10.) single-organism behavior
11.) cellular ion homeostasis
12.) rhythmic process

With the aim of reproducing the knowledge listed above, for every cluster in Figure 12.9, ORA was performed using the R package ORA [Lippmann et al., 2016]. The resulting p-values were filtered via ABC analysis, and thereafter, only group A was considered for interpretation (see chapter 9 for further details).

12.2.2 Knowledge Acquisition in Clusters of Pain Genes

DBS identified eight clusters[75] of genes (Figure 12.9). For each cluster, an ORA was performed. In contrast to the standard approach, in which the Bonferroni correction [Perneger, 1998] is

[75] After inspection of the functional areas in the eight ORA results, the eight clusters could be reduced to six (for details, see Tab. 2

often used, here, the p-values of the GO terms in the ORA results were filtered via ABC analysis [Ultsch/Lötsch, 2015]. The Bonferroni correction reduces the alpha error of significance, but it may cause valid results to be disregarded because the beta error simultaneously increases (for extensive discussions, see [Button et al., 2013; Nuzzo, 2014; Perneger, 1998]. Here, it is argued that in the special case of ORA, the p-values also represent the effect strength. Therefore, the adjustments to the significance threshold made by the Bonferroni correction are unnecessary. In contrast to the standard approach, ABC analysis was used to identify the most important GO terms as those assigned to group A, which had the highest effect strength. After the reduction of the directed acyclic graph (DAG) using this approach, the functional areas identified in [Lötsch et al., 2013] and [Ultsch et al., 2016b] were found to be associated with three of the classes (Table 12.2).

Considering the prior knowledge regarding pain functions and pain chronification, the following clusters could be combined: cluster 1 and cluster 3 were combined to class C1*, and cluster 2 and cluster 4 were combined into class C2*, because they showed similar functions and were separated only by low borders in the topographic map with hypsometric tints (Figure 12.9). Hence, it was possible to identify five classes with different semantic characterizations, plus one class of outliers (Tab. 2). Class C1* predominantly describes the pain functions of cells and reproduces knowledge presented in section 11.2.1. The main class (C2*) describes the molecular transport and signaling of pain, also reproducing prior knowledge about the pain genes. class C5 represents the downregulation of metabolic processes and the upregulation of the creatine metabolic process, which is a new discovery enabled by the DBS clustering. Class C6 describes outliers that are not relevant to the ORA-based DAG — these outliers are surrounded by very large hills in Figure 12.9. Class C7 characterizes the response and regulation systems as well as the upregulation of the phosphorus metabolic process, effectively reproducing the results of [Lötsch et al., 2013] and [Ultsch et al., 2016b]. The final class, C8, could represent hematopoietic stem cell differentiation. In summary, these clusters reproduce the previously identified functions of pain genes as described in section 11.2.1. In addition, new insights can also be found from class C5 and perhaps class C8.

Figure 12.11: The biological process of pain with the twelve functions of pain genes [Lötsch et al., 2013].

Table 12.2: Semantic characterization of the eight clusters of pain genes and the connections to prior knowledge. Downregulation is indicated as underlined, and new functional areas [Ultsch/Lötsch, 2014] are indicated in italics. The following clusters in Figure 12.9 were combined with the aid of prior knowledge: C1 and C3 were combined into C1*, and C2 and C4 were combined into C2*.

ORA Parameters	**Class.**	**No. of Genes**	**Semantic Meaning as Defined by GO Terms in ORA**	**Semantic Characterization**
RAW and Bonferroni, minimum number of genes=10	C1*	147	single-organism cellular process cell communication cellular response to stimulus localization cell death cellular ion homeostasis nervous system development single-organism behavior rhythmic process intracellular signal transduction anatomical structure development cellular ion homeostasis	Pain functions of cells
RAW and Bonferroni, minimum number of genes=10	C2*	232	synaptic transmission ion transport G protein-coupled receptor signaling pathway *transmembrane transport*	Molecular transport and signaling
RAW	C5	53	*creatine metabolic process* *metabolic process*	Downregulation of metabolic processes and upregulation of the creatine metabolic process
RAW	C6	21	None	Outliers
RAW and Bonferroni, minimum number of genes=2	C7	58	response to stress phosphorus metabolic process behavior positive regulation of biological process response to organic substance response to wounding regulation of localization regulation of system process	Response and regulation systems as well as upregulation of the phosphorus metabolic process
RAW	C8	17	*hematopoietic stem cell differentiation*	Hematopoietic stem cell differentiation