Similarity Measures for Clustering SNP and Epidemiological Data

Silvia Selinski

SFB 475, Fachbereich Statistik, Universität Dortmund

and the GENICA Network

Interdisciplinary Study Group on Gene Environment Interaction and Breast Cancer in Germany,

represented by C. Justenhoven (Stuttgart), H. Brauch (Stuttgart), S. Rabstein (Bochum), B. Pesch (Bochum), V. Harth (Bonn/Bochum), U. Hamann (Heidelberg), T. Brüning (Bochum), Y. Ko (Bonn)

Abstract

The issue of suitable similarity measures for a joint consideration of so called SNP data and epidemiological variables arises from the GENICA (Interdisciplinary Study Group on Gene Environment Interaction and Breast Cancer in Germany) case-control study of sporadic breast cancer. The GENICA study aims to investigate the influence and interaction of single nucleotide polymorphic (SNP) loci and exogenous risk factors. A single nucleotide polymorphism is a point mutation that is present in at least 1 % of a population. SNPs are the most common form of human genetic variations.

In particular, we consider 43 SNP loci in genes involved in the metabolism of hormones, xenobiotics and drugs as well as in the repair of DNA.

Assuming that these single nucleotide changes may lead, for instance, to altered enzymes or to a reduced or enhanced amount of the original enzymes – with each alteration alone having minor effects – the aim is to detect combinations of SNPs that under certain environmental conditions increase the risk of sporadic breast cancer.

The search for patterns in the present data set may be performed by a variety of clustering and classification approaches. I consider here the problem of suitable

measures of proximity of two variables or subjects as an indispensable basis for a further cluster analysis. In the present data situation these measures have to be able to handle different numbers and meaning of categories of nominal scaled data as well as data of different scales.

Generally, clustering approaches are a useful tool to detect structures and to generate hypothesis about potential relationships in complex data situations. Searching for patterns in the data there are two possible objectives: the identification of groups of similar objects or subjects or the identification of groups of similar variables within the whole or within subpopulations. The different objectives imply different requirements on the measures of similarity. Comparing the individual genetic profiles as well as comparing the genetic information across subpopulations I discuss possible choices of similarity measures suitable for genetic and epidemiological data, in particular, measures based on the χ^2-statistic, Flexible Matching Coefficients and combinations of similarity measures.

KEY WORDS: GENICA, single nucleotide polymorphism (SNP), sporadic breast cancer, similarity, cluster analysis, Flexible Matching Coefficient, Pearson's Corrected Coefficient of Contingency, mixed similarity coefficient

1. Introduction

The issue of the appropriate choice of measures of proximity arises from the GENICA (Interdisciplinary Study Group on Gene Environment Interaction and Breast Cancer in Germany) case-control study of sporadic breast cancer. In Germany almost 50 000 women develop breast cancer each year, that are 7 to 10 % of all women developing this disease during their life-time. Though genetic factors have been discovered for hereditary breast cancer – variations of the genes BRCA1 and BRCA2 in about 3 % of all cases – for the majority of the breast cancer cases such understanding of the genetic mechanisms and potential interactions with exogenous risk factors remains unclear. It is supposed that combinations of a number of low penetrant susceptibility genes may augment the risk of breast cancer in presence of certain exogenous risk factors. One of these factors seems to be the long term use of the Hormone Replacement Therapy as it was confirmed by the British Million Woman Study (Beral, 2003). Identification of interacting sequence variants and exogenous risk factors which affect the individual susceptibility is a major challenge for understanding the mechanisms contributing to the development of sporadic breast cancer (see also Garte, 2001).

The GENICA study aims to investigate these supposed genetic and gene-environment interactions associated with sporadic breast cancer. With respect to the genetic data the GENICA study group considers in particular single nucleotide polymorphisms (SNPs) – the most common genetic variation – in genes involved, for instance, in the metabolism of hormones and of xenobiotics and drugs, as well as of signal transductors. Besides the genetic traits the GENICA study considers a number of epidemiological variables which encompass a broad range of potential risk and beneficial factors such as age, physical activity, hormone use etc.

The search for patterns in the present data set may be performed by a variety of clustering and classification approaches. I consider here the problem of suitable measures of proximity of two variables or subjects as an indispensable basis for a further cluster analysis. This is also important for several classification approaches such as *k* Nearest Neighbours for non-metric dissimilarity measures (Zhang & Srihari, 2002).

The appropriate choice of measures of similarity requires a consideration of the concept of similarity and dissimilarity in the context of the particular data situation. That means to ascertain that candidate measures correspond to the scale of the data, that they are able to handle the specific difficulties of the data set, and, moreover, that the chosen measures reflect our believe about the nature of our data. For instance, measures based on the χ^2-statistic regard objects as dissimilar if they are independent and similar if they are dependent in the sense that certain combinations of categories occur more often than expected under the hypothesis of independence. These prominent combinations need not to be those of equal entries for each of the two objects. The latter is the concept of similarity underlying the matching coefficients.

This group of measures is particularly suitable for a comparison of SNP data and for a comparison of subjects, especially the Flexible Matching Coefficients, which may account for biological background knowledge and for the problem of the huge amount of homozygous reference sequences (Selinski & Ickstadt, 2005). This is a typical problem of SNP data and leads to a masking effect of the jointly occurring homozygous references with respect to the comparably small fraction of other jointly occurring genotypes and of dissimilar genotypes.

A further difficulty of the present data set is the diversity of the considered exogenous factors. Though the genetic data owns mainly the same structure – three categories with each category having a similar meaning – this is obviously not true for epidemiological variables such as smoking habits and family history of cancer. First of all we have to account for the different scale of the data. Moreover, within the categorial variables it is usually not possible to consider certain categories as similar. Furthermore, different concepts of similarity might be appropriate for subgroups of variables.

Hence, the question is, how to assign a numerical value measuring the proximity – similarity or dissimilarity – of two SNP loci, of two variables of different numbers of categories and different meaning or of different scale and how to measure the proximity of the genetic and epidemiologic profiles of two persons based on such a set of variables?

After a short introduction to the data the third section considers measures of proximity in general and particular measures for different scales and concepts of

similarity. Flexible Matching-Coefficients and combinations of measures are presented followed by some first results and conclusions in section four.

2. Data

Single Nucleotide Polymorphisms

SNP data are qualitative data providing information about the genotype at a specific locus of a gene. To be more precisely, a SNP (single nucleotide polymorphism) is a point mutation present in at least 1 % of a population. A point mutation is a substitution of one base pair or a deletion, which means, the respective base pair is missing, or an addition of one base pair. Though several different sequence variants may occur at each considered locus usually one specific variant of the most common sequence is found, an exchange from adenine (A) to guanine (G), for instance. Thus, information is basically given in form of categories denoting the combinations of base pairs for the two chromosomes, e.g. A/A, A/G, G/G, if the most frequent variant is adenine and the single nucleotide polymorphism is an exchange from adenine to guanine.

The result of such a variation of one base pair may be, for instance, a change of one amino acid in the amino acid chain of an enzyme or the switch from an amino acid coding triplet to a stop codon leading to a shortened amino acid chain. So, what we have to compare with respect to their similarity are present or absent alterations of certain base pairs of the DNA and the consequences of the altered genetic code with respect to the related metabolic processes and with respect to certain exogenous factors (see Selinski & Ickstadt, 2005, for more details).

GENICA case-control study

The present data set consists of a selection of SNP loci and epidemiological variables of the GENICA study of sporadic breast cancer. The GENICA study is a population-based age-matched case-control study assessing genotypes of over 100 SNP loci and exogenous risk factors of the reproductive history, hormone use, life style factors,

occupational history, family history of cancer, etc. of > 1000 cases and > 1000 healthy controls.

The GENICA network is a cooperation between researchers from the Research Institute for Occupational Medicine of the Institutions for Satutory Accident Insurance and Prevention (BGFA) in Bochum, the Dr.-Margarete-Fischer-Bosch Institute for Clinical Pharmacology (IKP) in Stuttgart, the German Cancer Research Center (DKFZ) in Heidelberg, the Medical Polyclinic at the University of Bonn, and the Institute for Occupational Physiology at the University of Dortmund (IfADo). The study is part of the German Human Genome Project (DHGP).

Actually the available data set comprises 43 SNP loci of 610 cases of sporadic breast cancer and of 650 age-matched healthy controls from the first phase of recruitment.

The main part of the SNP data are given in form of both detected bases at a specific locus, specifying the reference base and the variant, and are transformed to denote the single or double absence of the reference base pair at a defined point of a certain gene. In particular, we denote 0 as the homozygous reference sequence (reference/reference, no SNP), 1 as the heterozygous genotype (reference/variant, 1 SNP) and 2 as the homozygous variant sequence (variant/variant, 2 SNPs).

Furthermore, we know which loci belong to the same gene and to which pathways the genes belong to. Additionally, we know for most loci if they are located in a coding or in a non-coding region and in case of the coding SNP loci if they cause a change in the amino acid chain. Several genes are observed at more than one SNP locus and the pathway information is given for all genes (Selinski & Ickstadt, 2005). Pathway means the field where a gene-product plays a role within the human metabolism, e.g. the pathway of xenobiotics and drug metabolism. Note, that a gene may participate in more than one pathway.

Additionally, a selection of 49 categorial and 8 quantitative epidemiological variables is considered.

3. Methods

Searching for patterns in the data there are two possible objectives: a comparison of variables and a comparison of subjects. In the first case the aim is to detect major differences in the clustering of two variables between cases and controls as well as a general structure of genetic and or exogenous variables. A different point of view is the comparison of subjects. Here the objective is to find high and low risk groups with similar profiles of genetic variables and exogenous risk factors. Depending on the different objectives we have to define a measure of proximity suitable for the hypothised concept of similarity and the scale of the data.

A detailed introduction into the special issue of measures for SNP data is given by Selinski & Ickstadt (2005).

3.1 Concepts of proximity

Similarity may be considered in terms of *agreement* or in terms of *dependence*.

Agreement means to consider two variables as similar if the majority of the subjects own a combination of similar traits. Two variables would be considered as dissimilar if the majority of subjects have a combination of dissimilar traits. Similar traits may be equal categories in case of categorial data or the common occurrence of high or low values in case of quantitative data. Matching coefficients and measures of correlation, for instance, would correspond to this concept of similarity. Application of this concept requires

i. equal numbers of categories and assignment of similar categories or
ii. at least ordinal scale with a sufficient number of categories and similar meaning of high and low values

The concept of *dependence* encompasses the first in so far as a frequent occurrence of similar traits would also be regarded as similarity. But it also allows, in case of categorial data, generally for further combinations of – perhaps a priori judged as dissimilar – to contribute to the label ‘similar’ for two variables or subjects if they occur more frequent than expected. So, dependence would be regarded as similarity and independence as dissimilarity. This concept is represented, for instance, by squared correlation coefficients in case of quantitative or ordinal scaled data and measures based on the χ^2-statistics in case of categorial data.

Focusing on the similarity of the genetic and epidemiological variables the basic questions are: What does similarity of two SNP loci mean, what does similarity of variables of different interpretation and scale mean and how to measure it?

With respect to SNP loci we may regard the common occurrence or absence of sequence alterations as similarity – and apply matching coefficients – or we may regard dependence as similarity – and apply measures based on the χ^2-statistic. For a comparison of epidemiologic variables and for a joint comparison of genetic and epidemiologic variables we have to consider the following cases:

i. All variables are categorial but with different numbers of categories. Equally denoted categories may have a different interpretation.
ii. Most variables are categorial, some of them are ordinal scaled, the remaining variables are quantitative. Equally denoted categories may have a different interpretation.

In the first case we can use measures based on the concept of dependence suitable for categorial data, e.g. Pearson's Corrected Coefficient of Contingency, for the comparison of variables. For the comparison of subject we may additionally use Matching Coefficients. In the second case there are two possibilities. The quantitative variables may be transformed to categorial variables with a sufficient low number of categories to avoid empty cells. Hence, we can proceed as in the categorial case. The second option is to use different measures of similarity for the different scales and to combine them to a coefficient for different scales.

Focussing on the comparison of objects or subjects – the observed persons in this case – means to assess the similarity of each trait of the two subjects separately and to draw conclusions about the overall similarity of the considered genetic and epidemiologic profiles. Generally, two subjects can be considered as similar if they share mainly similar traits. They are dissimilar if most considered variables show dissimilar combinations of traits. Thus similarity means here *accordance* or *agreement*. The concept of *dependence* is less adequate. Imagine that the genotypes of two persons are compared by means of a measure based on the χ^2-statistic. Then they would be regarded as similar if the observed cell counts deviate from the expected ones. This means not necessarily that they share the same genotype at most loci. We would obtain the same result if they share the same genotype at notably few

loci - in contrast to our believe about similarity in this situation. So, in this particular situation measures based on the concept of agreement should be preferred to those based on dependence.

A general problem of SNP data is the huge amount of common occurrence of homozygous reference types which is supposed to mask the relevant information of combinations of genetic alterations. Especially the Flexible Matching Coefficients introduced in section 3.3 are able to handle this specific problem of such data sets.

3.2 Similarity and distance

Measures of similarity or distance may be defined as functions of variables or as functions of objects or subjects. We introduce here functions of variables. For the corresponding notations of the functions of objects replace $S : V \times V \to IR$, with V being the set of variables by $S : O \times O \to IR$, with O being the set of objects.

DEFINITION 1. Similarity

Let $O = \{O_1, \ldots, O_n\}$ be a set of n objects observed at a set of m variables $V = \{V_1, ..., V_m\}$. Then a measure of similarity of two variables $V_k \in V$ and $V_l \in V$, is given by $S : V \times V \to IR$ with

(A1)	$S(V_k, V_l) > S(V_k, V_m)$,	$\forall\ V_k, V_l, V_m \in V$, with V_k being more similar to V_l than to V_m and $V_l \neq V_m$	comparability
(A2)	$S(V_k, V_l) = S(V_l, V_k)$,	$\forall\ V_k, V_l \in V$	symmetry
(A3)	$S(V_k, V_k) \geq S(V_k, V_l)$,	$\forall\ V_k, V_l \in V$	natural order

REMARK 1. Restriction to [0,1]

Often it is useful to assume that $S \in [0,1]$, i.e.,

(A4)	$S(V_k, V_l) \geq 0$,	$\forall\ V_k, V_l \in V$	positivity
(A5)	$S(V_k, V_k) = 1$,	$\forall\ V_k, \in V$	normality

Measures of distance or dissimilarity can be defined similarly.

DEFINITION 2. Distance

Let $O = \{O_1, \ldots, O_n\}$ be a set a set of n objects observed at a set of m variables $V = \{V_1, \ldots, V_m\}$. Then a measure of distance of two variables $V_k \in V$ and $V_l \in V$, is given by $D : V \times V \to IR$ with

(B1)	$D(V_k, V_l) > D(V_k, V_m)$,	$\forall\ V_k, V_l, V_m \in V$, with V_k being more dissimilar to V_l than to V_m and $V_l \neq V_m$	comparability
(B2)	$D(V_k, V_l) = D(V_l, V_k)$,	$\forall\ V_k, V_l \in V$	symmetry
(B3)	$D(V_k, V_k) \leq D(V_k, V_l)$,	$\forall\ V_k, V_l \in V$.	natural order

REMARK 3. Restriction to [0,1]

Often it is useful to assume that $D \in [0,1]$, i.e.,

(B4)	$D(V_k, V_l) \leq 1$,	$\forall\ V_k, V_l \in V$	positivity
(B5)	$D(V_k, V_k) = 0$,	$\forall\ V_k \in V$.	normality

REMARK 4. Metric

If D satisfies (B2),

(B6)	$D(V_k, V_l) = 0$,	if and only if $k = l$, $\forall\ V_k, V_l \in V$	normality
(B7)	$D(V_k, V_l) + D(V_l, V_m) \geq D(V_k, V_m)$,	$\forall\ V_k, V_l, V_m \in V$ and $V_l \neq V_m$	triangle inequality

then D is a metric.

Note, that (B6) is a stronger assumption than (B5). Furthermore, D is not restricted to [0, 1].

In practice, the interest is focussed more on distances, especially on metric measures of distances. If $S \in [0, 1]$ then $D = 1 - S$ otherwise S can be converted into a distance as follows:

TRANSFORMATION 1.

Let S be a similarity measure satisfying (A1)-(A3) and let $\min S(V_k, V_l) < 0$. Then the transformation

(T1) $$D(V_{k'}, V_{l'}) = 1 - \frac{S^*(V_{k'}, V_{l'})}{\max S^*(V_k, V_l)}, \qquad \forall\ V_{k'}, V_{l'} \in V \text{ and } \forall V_k, V_l \in V,$$

where $S^*(V_{k'}, V_{l'}) = S(V_{k'}, V_{l'}) + |\min S(V_k, V_l)|$, $\forall\ V_{k'}, V_{l'} \in V$ and $\forall V_k, V_l \in V$,

yields the corresponding measure of distance $D : V \times V \to [0,1]$.

If S also satisfies (A4) the transformation from S to S^* can be skipped and (T1) can be performed directly with S.

If S in addition satisfies (A5) the transformation

(T2) $$D(V_k, V_l) = 1 - S(V_k, V_l), \qquad \forall\ V_k, V_l \in V,$$

yields the corresponding measure of distance $D : V \times V \to [0,1]$.

3.3 Measures of proximity

Choosing appropriate measures of proximity for a particular problem does not only mean to regard the nature of similarity and dissimilarity but also to consider the scale of the data and special characteristics of the data set. This section considers the different scales of data and gives an overview over the corresponding measures of proximity focussing on the particular situation of SNP and epidemiologic data.

Nominal scale

Considering the similarity of nominal scaled data in terms of agreement the corresponding measures of agreement relate the numbers of pairs of similar traits to the number of pairs of dissimilar traits. There is a plethora of similarity measures based on this concept. We concentrate here on Flexible Matching Coefficients that encompass most of the common matching coefficients (Selinski & Ickstadt, 2005). For further matching coefficients that may not be derived from the following Definition 3 see, for instance, Anderberg (1973), Cox & Cox (2001), Steinhausen & Langer (1977).

Measures of dependence are usually based on the χ^2-statistic and differ in their way of handling the dependence of the χ^2-statistic on the table size.

Measures of agreement

Consider the case of V_k and V_l with categories $k, l = 0, 1, \ldots, p$ being two variables that should be compared with respect to their similarity. The case of O_k and O_l is analogous. It is reasonable to assume that the matching categories are all combinations i-j with $i = j$, $i, j = 0, 1, \ldots, p$.

In the particular situation of SNP data this means that we compare either loci or persons with the matching combinations

0-0 homozygous reference- homozygous reference,

1-1 heterozygous-heterozygous and

2-2 homozygous variant- homozygous variant.

where extensions to further combinations are possible.

So, let V_k and V_l with categories $i, j = 0, 1, \ldots, p$ being two variables and let m'_{ij} as given in Table 1.

Table 1. *Contingency table of V_k and V_l.*

V_k \ V_l	0	1	2	…	p
0	m_{00}	m'_{01}	m'_{02}	…	m'_{0p}
1	m'_{10}	m_{11}	m'_{12}	…	m'_{1p}
2	m'_{20}	m'_{21}	m_{22}	…	m'_{2p}
…	…	…	…	…	…
p	m'_{p0}	m'_{p1}	m'_{p2}	…	m_{pp}

For convenience and to assure the symmetry of the corresponding similarity matrix for all variables or subjects the indices kl and lk are pooled together to one index kl, $k \le l$. Note that $m_{kl} = m'_{kl} + m'_{lk}$, $\forall k, l = 1, \ldots, p$, $k \le l$, is the sum over all numbers of categories k and l.

DEFINITION 3. Flexible Matching Coefficient

Let $O = \{O_1, \ldots, O_n\}$ be a set n objects observed at a set of m variables $V = \{V_1, \ldots, V_m\}$. Then $S^{flex\text{-}IJ,\lambda,\delta}$: $V \times V \to IR$, and $S^{flex\text{-}IJ,\lambda,\delta}$: $O \times O \to IR$, respectively, is given by

$$S^{flex-IJ,\lambda,\delta} := \frac{\Lambda}{\Lambda + \Delta}, \qquad (1)$$

with $\Lambda := \sum_{i \in I} \lambda_i m_i$, $\Delta := \sum_{j \in J} \delta_j m_j$,

$I = \{i=kl, k \le l, k, l = 0, 1, \ldots, p | \text{all combinations of category } k \text{ and } l \text{ are similar}\}$,

$J = \{j=kl, k \le l, k, l = 0, 1, \ldots, p | \text{all combinations of category } k \text{ and } l \text{ are dissimilar}\}$.

We denote by λ the vector of weights λ_i, $i \in I$, of the matches and by δ the vector of weights δ_j, $j \in J$, of the mismatches. Furthermore, $\lambda_i \ge 0, \forall i \in I$, $\sum_{i \in I} \lambda_i > 0$, $\delta_j \ge 0, \forall j \in J$, $\sum_{j \in J} \delta_j > 0$, and $m_i \ge 0, \forall i \in I$, $m_j \ge 0, \forall j \in J$, $\sum_{i \in I} m_i + \sum_{j \in J} m_j > 0$ with m_i denoting the number of entries of all combinations of matching categories contributing to i and m_j denoting the number of entries of all combinations of dissimilar categories contributing to j. In particular, $m_{kl} = m'_{kl} + m'_{lk}$ is the sum of the number of (k, l) and (l, k) pairs.

REMARK 5. Measure of Similarity

$S^{flex-IJ,\lambda,\delta} = \frac{\Lambda}{\Lambda + \Delta}$ is a measure of similarity satisfying (A1)-(A5).

PROOF: see Selinski & Ickstadt (2005).

REMARK 6. Special cases for SNP data

For the comparison of SNP data as described in section 2 the following special cases of (1) can be applied:

i. With $I = \{0, 1, 2\}$ and $J = \{02, 01, 12\}$ we obtain

$$S^{flex-ii,\lambda,\delta} = \frac{\lambda_2 m_{22} + \lambda_1 m_{11} + \lambda_0 m_{00}}{\lambda_2 m_{22} + \lambda_1 m_{11} + \lambda_0 m_{00} + \delta_{02}(m_{02} + m_{20}) + \delta_{01}(m_{01} + m_{10}) + \delta_{12}(m_{12} + m_{21})} \qquad (2)$$

$\lambda_i \ge 0$, $i = 0, 1, 2$, $\delta_j \ge 0, j = 02, 01, 12$, $\sum_i \lambda_i > 0$, $\sum_j \delta_j > 0$.

It might be reasonable to assume that $\lambda_2 \ge \lambda_1 \ge \lambda_0 \ge 0$ stressing the importance of the common occurrence of homozygous variants, that $\delta_{02} \ge \delta_{01} > 0$ and $\delta_{02} \ge \delta_{12} > 0$ so that homozygous variants and references are set to be most dissimilar.

ii. With $I = \{0, 1, 2, 12\}$ and $J = \{02, 01\}$ we obtain

$$S^{flex-12,\lambda,\delta} = \frac{\lambda_2 m_{22} + \lambda_1 m_{11} + \lambda_0 m_{00} + \lambda_{12}(m_{12} + m_{21})}{\lambda_2 m_{22} + \lambda_1 m_{11} + \lambda_0 m_{00} + \lambda_{12}(m_{12} + m_{21}) + \delta_{02}(m_{02} + m_{20}) + \delta_{01}(m_{01} + m_{10})} \qquad (3)$$

$\lambda_i \geq 0$, $i = 0, 1, 2, 12$, $\delta_j \geq 0$, $j = 02, 01$, $\sum_i \lambda_i > 0$, $\sum_j \delta_j > 0$.

It might be reasonable to assume that $\lambda_2 \geq \lambda_1 \geq \lambda_0 \geq 0$, $\lambda_2 \geq \lambda_1 \geq \lambda_{12} \geq 0$ and $\delta_{02} \geq \delta_{01} > 0$.

iii. With $I = \{0, 1, 2, 01\}$ and $J = \{02, 12\}$ we obtain

$$S^{flex-01,\lambda,\delta} = \frac{\lambda_2 m_{22} + \lambda_1 m_{11} + \lambda_0 m_{00} + \lambda_{01}(m_{01} + m_{10})}{\lambda_2 m_{22} + \lambda_1 m_{11} + \lambda_0 m_{00} + \lambda_{01}(m_{01} + m_{10}) + \delta_{02}(m_{02} + m_{20}) + \delta_{12}(m_{12} + m_{21})} \quad (4)$$

with $\lambda_i \geq 0$, $i = 0, 1, 2, 01$, $\delta_j \geq 0$, $j = 02, 12$, $\sum_i \lambda_i > 0$, $\sum_j \delta_j > 0$.

It might be reasonable to assume that $\lambda_2 \geq \lambda_1 \geq \lambda_0 \geq 0$, $\lambda_2 \geq \lambda_1 \geq \lambda_{01} \geq 0$ and $\delta_{02} \geq \delta_{12} > 0$.

For the properties of $S^{flex-IJ,\lambda,\delta}$ and the relationship between Flexible and conventional Matching Coefficients, see Selinski & Ickstadt (2005).

The flexibility of the sets of similar and dissimilar indices enables an incorporation of biological assumptions about dominance. Equation (3) corresponds to the dominance of the variant sequence, Equation (4) to the dominance of the reference.

The Flexible Matching-Coefficients are especially suitable for the comparison of SNP data. For a comparison of subjects based on SNP and epidemiological data it is not possible to take advantage of the flexibility of the measure. The categories of the epidemiological variables are so different from each other and from the SNP data, that we cannot a priori judge certain categories as more or less important for the similarity of two subjects neither we can define other combinations than those of equal entries to contribute rather to the similarity than to the dissimilarity. So, in this case the we require $\lambda_i = \lambda_{i*}, \forall i, i* \in I$ and $\delta_j = \delta_{j*}, \forall j, j* \in J$.

In case of a comparison of genetic and epidemiological variables there is generally the problem that equally denoted categories do not necessarily have the same meaning. For instance, let X be a SNP locus and Y be categories of the numbers of mammograms in categories of 'never', '1-9' and '10+' by '0', '1' and '2', respectively. Hence, it is clear that the number of equal entries does not reveal anything about the similarity of these two variables. The concept of similarity that is more appropriate in this situation is the concept of dependence as similarity as introduced in the next section.

Measures of dependence

In case of nominal scaled data most measures based on the concept of dependence are functions of the χ^2-statistic and handle the problem of the dependence of this statistic on the table size differentially (Anderberg, 1973, Hartung, 1991) as, for instance, Pearson's Corrected Coefficient of Contingency

$$S_{PC} = \sqrt{\frac{\min(p,q)}{\min(p,q)-1}} \cdot \sqrt{\frac{\chi^2}{\chi^2+m}}, \qquad (5)$$

where p and q are the numbers of categories of the variables or objects which should be compared, $m = \sum_{i=1}^{p}\sum_{j=1}^{p} m_{ij}$ is the number of observations contributing to χ^2, $0 \le C = \sqrt{\frac{\chi^2}{\chi^2+m}} \le \sqrt{\frac{\min(p,q)-1}{\min(p,q)}} < 1$ is Pearson's Contingency Coefficient and the factor $\sqrt{\frac{\min(p,q)}{\min(p,q)-1}}$ is used to eliminate the dependence of C from the table size.

A further member of this class of measures, Cramèr's C, is considered in Müller *et al.* (2005).

Pearson's Corrected Coefficient of Contingency seems to be a useful tool to compare categorial variables. In particular, S_{PC} allows for different numbers of categories and we do not have to specify similar categories of two variables previous to an analysis. So, we are able to compare variables which are so different from each other that we don't have an idea which might be similar categories and which ones are dissimilar, e.g. the genotypes at a SNP locus in a gene coding for NAT2 and the number of children recorded in categories '0', '1', '2', '3-4', '>4'. Thus, S_{PC} seems to be a useful tool for a comparison of genetic and exogenous risk factors.

Generally it is possible to use S_{PC} for the comparison of subjects. Note, that two objects would also be regarded as similar if they have notably few equal entries for the same variables. So the use of S_{PC} seems to be not the best choice to cluster subjects.

Ordinal scale

In case of ordinal scaled data we can assess the proximity of two variables or subjects using measures based on the concept of *correlation* or on the concept of *dependence*. The latter can be obtained from correlation coefficients by squaring them. Coefficients of correlation have to be suitable for ordinal scaled data, Spearman's rank correlation coefficient or Kendall's τ, for instance, and it would be reasonable to account for ties.

Considering proximity in terms of correlation means to regard a positive correlation as *similarity* and a negative correlation as *dissimilarity*. Correlation coefficients are restricted to [-1, 1], so transforming them into a measure of distance Transformation (T1) has to be applied, i.e. to obtain the corresponding measure of similarity from Kendall's τ

$$S_\tau = \frac{\tau_{corr} + 1}{2}, \tag{6}$$

where τ_{corr} denotes the correlation coefficient corrected for ties (Hollander & Wolfe, 1999). Considering correlation – positive or negative – as *similarity* and independence as *dissimilarity* suitable measures of proximity may easily be derived from correlation coefficients for ordinal data by using the square of these coefficients. Hence, the resulting measures of proximity are already standardised to [0, 1]. Note, that the applied coefficients of correlation should also be corrected for ties.

In the present case part of the epidemiological variables can be considered as ordinal scaled, categories of the years of oral contraceptive use, for instance. In case of SNP data it is possible to define an order in the determined genotypes in terms of the amount of the original gene dose: To interpret the homozygous reference type as double presence of the reference sequence (set to 2 or 1), the heterozygous type as single presence of the reference sequence (set to 1 or 0.5) and the homozygous variant type as absence of the reference sequence (set to 0).

Hence, coefficients of correlation may be used as a measure of similarity comparing subjects or variables and squared coefficients of correlation may be used additionally for a comparison of variables. The difficulty with this approach is that in case of SNP data we have only three possible categories for 1200 observations comparing the variables or three possible categories for over 60 observations for a comparison of subjects. This means that we have three tied groups that are quite large at the best.

So, this approach would be useful only in case of more than 3 categories that can be ordered and if the size of the tied groups is not too big. In particular this approach might be useful for a subset of epidemiological variables.

Quantitative data

Generally, there are three possibilities to consider the proximity of quantitative data: Association or rather correlation, dependence and a geometric interpretation of distance. The first concept leads to coefficients of correlation, Spearman's rank correlation coefficient or Kendall's τ, for instance, and is often applied in the analysis of gene expression data (see, for instance, Eisen *et al*., 1998). Similarity in terms of dependence and independence can be obtained applying squared correlation coefficients. The use of metric measures, Minkowski-r-metrics, for instance, is appropriate if the proximity or distance of two objects has a geometric interpretation, i.e. it is reasonable to require that the applied measure of proximity satisfies the triangle inequality (B7). Note that coefficients of correlation as well as metric measures are not necessarily restricted to [0, 1].

In the special case of SNP data it is clear that we actually don't have quantitative data. But some of the epidemiologic variables, such as the body mass index, are quantitative.

Mixtures of similarity coefficients

Since there seems to be no single measure of proximity that is suitable for all considered variables it is reasonable to apply the most suitable ones for parts of the variables and to form a joint similarity coefficient, for instance, as a weighted average.

The general idea is to split the set of m variables $V = \{V_1, \ldots, V_m\}$ into G subsets $V^1 = \{V_1, \ldots, V_{m_1}\}, \ldots, V^G = \{V_{m_1+\ldots+m_{G-1}+1}, \ldots, V_{m_1+\ldots+m_G}\}$ of variables so that within each subset a particular similarity measure can be applied.

Clustering of subjects

In case of a comparison of subjects this approach can be implemented quite easily. For instance, let

$g = 1$: all SNP loci with $p = 3$ categories and apply $S^{flex-IJ,\lambda,\delta}$,

$g = 2$: all SNP loci with $p \neq 3$ categories and all categorial epidemiological variables and apply $S^{flex-IJ,\lambda,\delta}$ with $I = \{ij,\ i = j\}$, $J = \{ij,\ i \neq j\}$, $\lambda_i = \lambda_{i*}, \forall i, i* \in I$ and $\delta_j = \delta_{j*}, \forall j, j* \in J$.

$g = 3$: all ordinal scaled variables with a sufficient number of categories and all quantitative variables and apply Kendall's τ corrected for ties and standardised according to Equation (6), denoted further by S_τ or its square, denoted by S_{τ^2} within V^3.

Calculate then the overall similarity matrix as a weighted average of the G subset similarity matrices.

There are generally two possibilities for weighting schemes: Equal weights for each entry of the respective subset similarity matrix or different weights for each entry of the respective subset matrix. The weights may be chosen, for instance, according to the assumed importance of the subsets of variables for the overall similarity, they may be chosen according to the number of variables in each subset or according to the numbers of observations contributing to each entry of the respective group matrix.

DEFINITION 4. Similarity Coefficient for clustering subjects

Let $O = \{O_1, \ldots, O_n\}$ be a set n objects observed at a set of m variables $V = \{V_1, \ldots, V_m\}$. Suppose that V can be divided into G subsets $V^1 = \{V_1, \ldots, V_{m_1}\}, \ldots,$ $V^G = \{V_{m_1+\ldots+m_{G-1}+1}, \ldots, V_{m_1+\ldots+m_G}\}$ of variables, with $m_1 + \ldots + m_G = m$, so that within each subset V^g, $g = 1, \ldots, G$, there exists a measure of similarity S^g which can be applied to measure the similarity of all subjects O_k and $O_l \in O$, $k, l = 1, \ldots, n$, based on this particular subset of variables V^g. Assume that S^g, $g = 1, \ldots, G$, satisfies (A1) – (A5). Let ω_{kl}^g be the weight for $S^g(O_k, O_l)$. Hence, $S^{mixed} : O \times O \to IR$ is given by

$$S^{mixed}(O_k, O_l) = \frac{\sum_{g=1}^{G} \omega_{kl}^g \cdot S^g(O_k, O_l)}{\sum_{g=1}^{G} \omega_{kl}^g}. \qquad (7)$$

The weights ω_{kl}^{g} may be chosen as the number of observations contributing to $S^{g}(O_k, O_l)$. This seems to be particularly useful in case of many missing observations. A further possibility is to choose ω_{kl}^{g} as the number of variables m_g in subset g or to determine a fixed weight ω^{g} to each subset according to the assumed importance of the respective subset. Hence Equation (7) can be simplified to

$$S^{mixed}(O_k, O_l) = \frac{\sum_{g=1}^{G} \omega^{g} \cdot S^{g}(O_k, O_l)}{\sum_{g=1}^{G} \omega^{g}}. \tag{8}$$

From Definition 4 it is obvious that S^{mixed} satisfies (A2) – (A5). To assure the comparability (A1) the weights have to reflect the importance of the groups of variables for the overall similarity. In case of equal weights for all pairs of objects it is reasonable to make sure, that the pairs of subjects do not differ remarkably from each other in the number of observations per group V^{g} available for a comparison. Imagine the situation where a pair of subjects O_k and O_l has notably few common observations in those groups V^{g} that are most important for a comparison of subjects and by chance the remaining observations indicate a high similarity. Hence, O_k and O_l may have a higher similarity coefficient as another pair O_k and O_m of subjects though there is no information about their similarity with respect to a remarkably high number of variables.

Clustering variables

Defining a coefficient for clustering variables analogous to Definition 4 is more difficult. Assume, that two variables belonging to the same group g may be compared using the respective measure of similarity S^{g}. Comparing two variables of different groups the similarity might either be set equal to zero or a further measure of similarity might be applied for a comparison. For instance, let

$g = 1$: all SNP loci with $p = 3$ categories and apply $S^{flex-IJ,\lambda,\delta}$ or S_{PC} within V^{1},

$g = 2$: all SNP loci with $p \neq 3$ categories and all categorial epidemiological variables and apply S_{PC} within V^{2},

$g = 3$: all ordinal scaled variables with a sufficient number of categories and all quantitative variables and apply S_{τ} or S_{τ^2} within V^{3}.

For a comparison two variables V_k and V_l belonging to different groups of variables V^g and V^{g^*}, respectively, apply

$S^{[1,2]} = S_{PC}$ in case of V^1 and V^2,

$S^{[1,3]} = 0$ or $S^{[1,3]} = S_{KW,}$ in case of V^1 and V^3,

$S^{[2,3]} = 0$ or $S^{[2,3]} = S_{KW,}$ in case of V^2 and V^3,

with $S_{KW} = 1\text{-}p_{Kruskal\text{-}Wallis,}$ $p_{Kruskal\text{-}Wallis}$ being the p-value of the Kruskal-Wallis test.

The resulting similarity matrix has the form of a block matrix with blocks of similarity coefficients for variables of the same group on the main diagonal and with blocks of zero or further coefficients of similarity for variables of different groups. A coefficient of similarity for a comparison of variables of different scales can then be defined as follows.

DEFINITION 5. Similarity Coefficient for clustering variables

Let $O = \{O_1, \ldots, O_n\}$ be a set n objects observed at a set of m variables $V = \{V_1, ..., V_m\}$. Suppose that V can be divided into G subsets $V^1 = \{V_1, \ldots, V_{m_1}\}, \ldots,$ $V^G = \{V_{m_1+\ldots+m_{G-1}+1}, \ldots, V_{m_1+\ldots+m_G}\}$ of variables, with $m_1 + \ldots + m_G = m$, so that within each subset V^g, $g = 1, \ldots, G$, there exists a measure of similarity $S^{[g,g]}$ which can be applied to measure the similarity of all variables V_k and $V_l \in V^g$, $k, l = 1, \ldots, m$. Let $S^{[g,g^*]}$, $g, g^* = 1, \ldots, G$, $g \neq g^*$, be a measure of similarity that can be applied to compare $V_k \in V^g$ and $V_l \in V^{g^*}$, $\forall\ V_k \in V^g$ and $V_l \in V^{g^*}$.

Assume that $S^{[g,g^*]}$, $g, g^* = 1, \ldots, G$,, satisfies (A1) – (A5). Let I_{V^g} be the indicator function.

Hence, $S^{block} : V \times V \to IR$ is given by

$$S^{block}(V_k, V_l) = \sum_{g=1}^{G} \sum_{g^*=1}^{G} S^{[g,g^*]}(V_k, V_l) \cdot I_{V^g}(V_k) \cdot I_{V^{g^*}}(V_l). \qquad (9)$$

From Definition 5 follows that S^{block} satisfies (A2) – (A5). Within each group of variables it is also obvious that S^{block} satisfies (A1). This is not necessarily the case if V_l and V_m belong to different groups of variables. So the choice of measures of similarity $S^{[g,g^*]}$ has to be handled with care.

4. Results

The calculation of the similarity matrices as well as the cluster analysis were performed using the software packages R.2.0.1 and R.1.8.0. For the cluster analysis the average linkage algorithm was applied (Kornrumpf, 1986, see also Sitterberg, 1978, and Ostermann & Degens, 1984, for properties of the average linkage algorithm).

A detailed comparison of the conventional matching coefficients and measures based on the χ^2-statistic is given in Müller *et al.* (2005) and Müller (2004), Flexible Matching Coefficients are considered by Selinski & Ickstadt (2005).

Results are shown for the clustering of SNP and categorial epidemiological variables for both: cases and controls. First, the dendrograms resulting from the application of S_{PC} are shown for the group of SNP loci and the categorial epidemiological variables separately as well as combined (Figures 1 to 6).

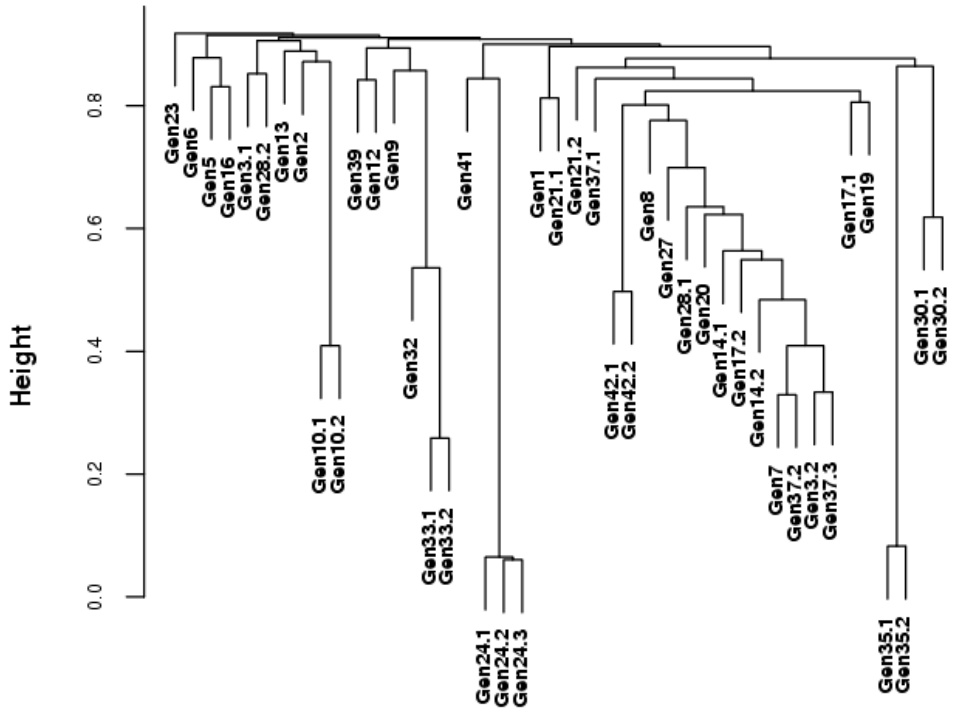

Figure 1. *Dendrogram of S_{PC} of the SNP loci of the control group.*

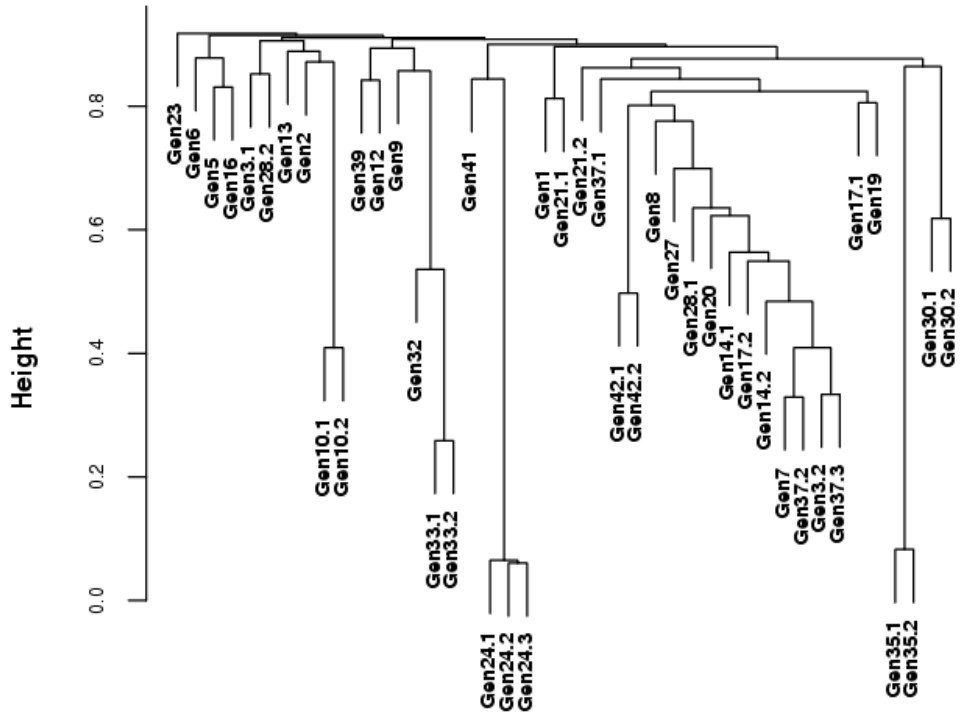

Figure 2. *Dendrogram of S_{PC} of the SNP loci of the case group.*

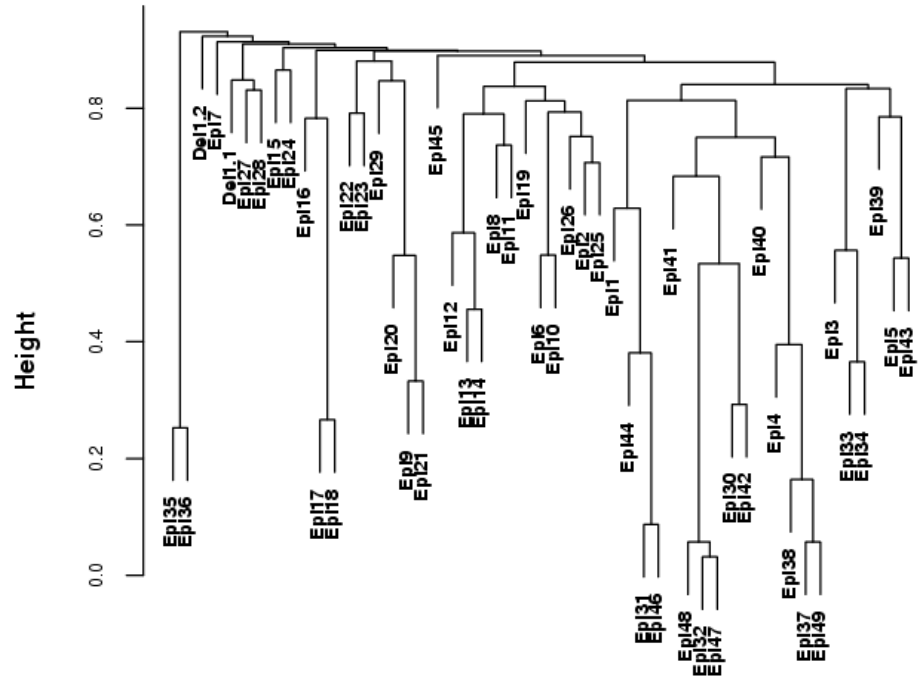

Figure 3. *Dendrogram of S_{PC} of the categorial epidemiological variables of the control group.*

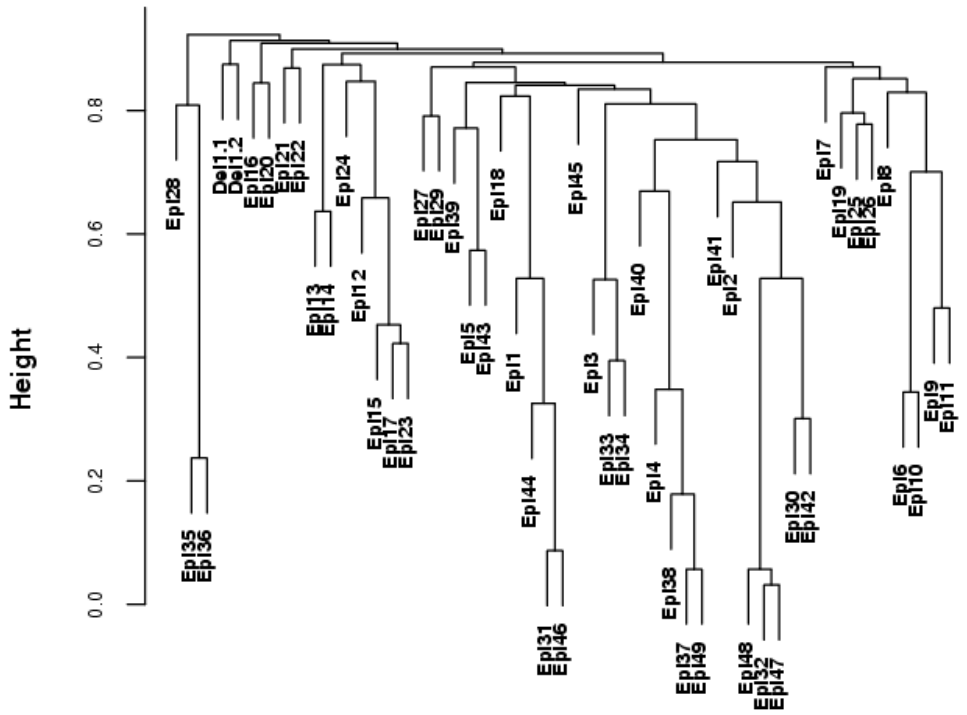

Figure 4. *Dendrogram of S_{PC} of the categorial epidemiological variables of the case group.*

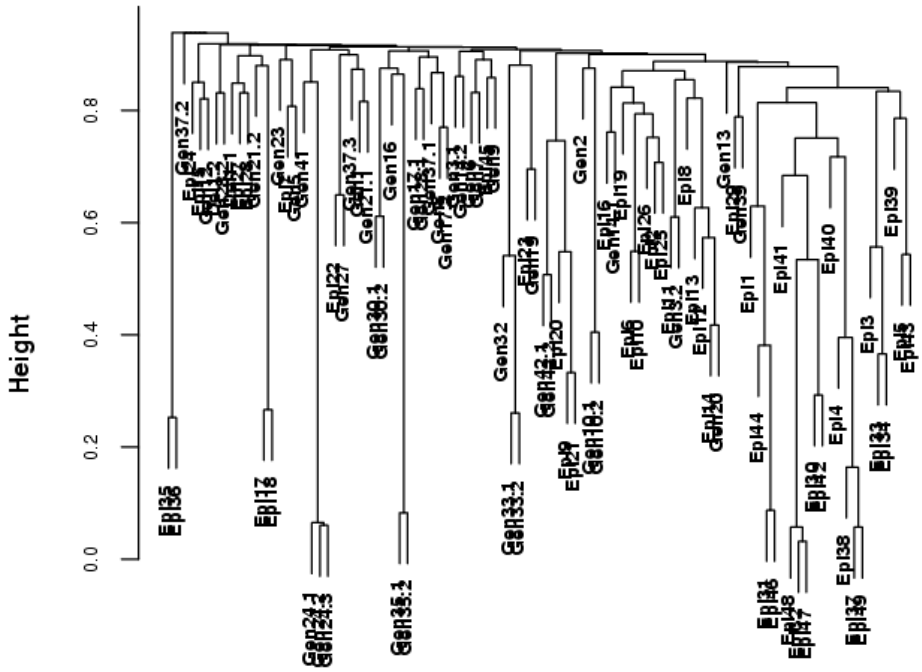

Figure 5. *Dendrogram of S_{PC} of the SNP loci and the categorial epidemiological variables of the control group.*

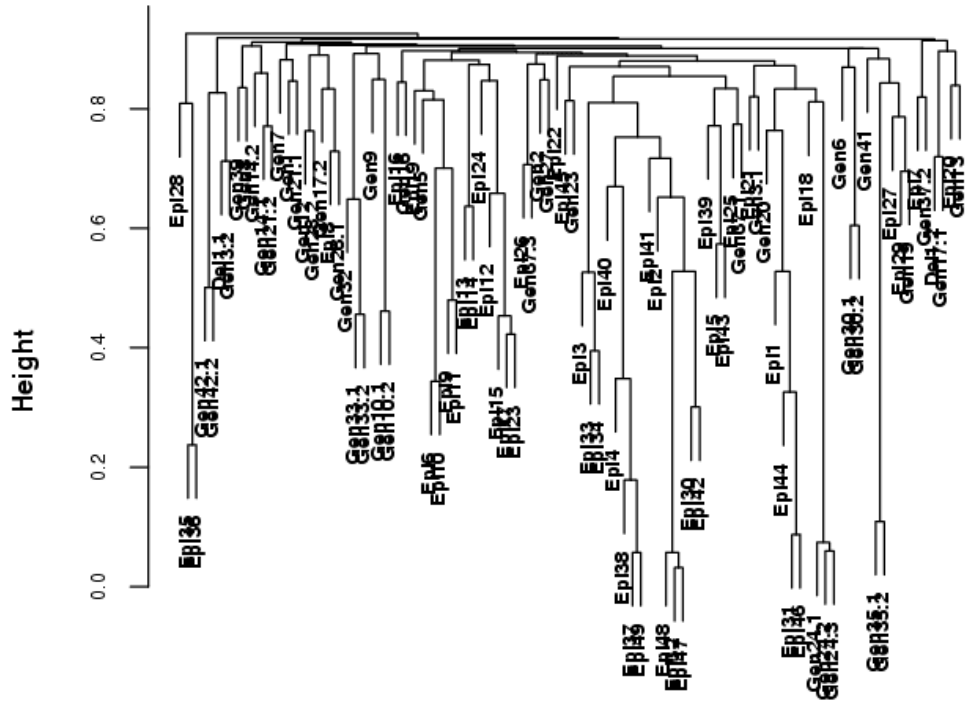

Figure 6. *Dendrogram of S_{PC} of the SNP loci and the categorial epidemiological variables of the case group.*

Figures 1 to 6 show that structures within subgroups of variables remain in a joint consideration. For instance, the cluster of the loci Gen41, Gen24.1, Gen24.2 and Gen24.3 in Figure 1 remains in Figure 5 as well as the cluster of Epi1, Epi3 – Epi5, Epi31 – Epi34, Epi37 – Epi44 and Epi46 – Epi49 in Figure 3. Single SNP loci and epidemiological variables can be found in common clusters. Here, main differences between cases and controls can be detected as, for example, the cluster Epi9, Epi20, Epi21, Gen42.1 and Gen42.2 in the control group (Figure 5) that cannot be found in the case group (Figure 6). In the case group Epi9 is clustered together with Epi6, Epi10, Epi11, Gen5 and Gen9, whereas Epi11 is clustered together with Epi8, Epi12 – Epi14, Gen3.2 and Gen20 in the control group, Gen5 is clustered together with Gen23 and Epi7. Gen9 is clustered together with Epi45, Epi6 and Epi10 are also joined together in the control group.

In the case group Epi20 is clustered together with Gen13 and Epi21 is clustered together with Gen3.1. The loci Gen42.1 and Gen42.2 are clustered together with the loci Gen3.2 and Del1.1 in the case group.

Applying S^{block} with g = 1 being the group of SNP loci, g = 2 being the group of categorial epidemiological variables and g = 3 being the group of the quantitative variables and using $S^{[1,1]} = S^{flex-IJ,\lambda,\delta}$, $S^{[2,2]} = S_{PC}$, $S^{[3,3]} = S_{\tau^2}$ as well as $S^{[3,3]} = S_{\tau}$, $S^{[1,2]} = S_{PC}$ and $S^{[1,3]} = S^{[2,3]} = S_{KW}$ as similarity coefficients separately for cases and controls results in Figures 7 to 14. Three different specifications for $S^{flex-IJ,\lambda,\delta}$ are used:

i. $I = \{2, 1, 0\}$, $J = \{12, 01, 02\}$, $\lambda = (2, 1, 1/3)$, $\delta = (2/3, 1, 2)$,

ii. $I = \{2, 1, 0, 12\}$, $J = \{01, 02\}$, $\lambda = (2, 1, 1/3, 2/3)$, $\delta = (1, 2)$,

iii. $I = \{2, 1, 0, 01\}$, $J = \{12, 02\}$, $\lambda = (2, 1, 1/3, 2/3)$, $\delta = (1, 2)$.

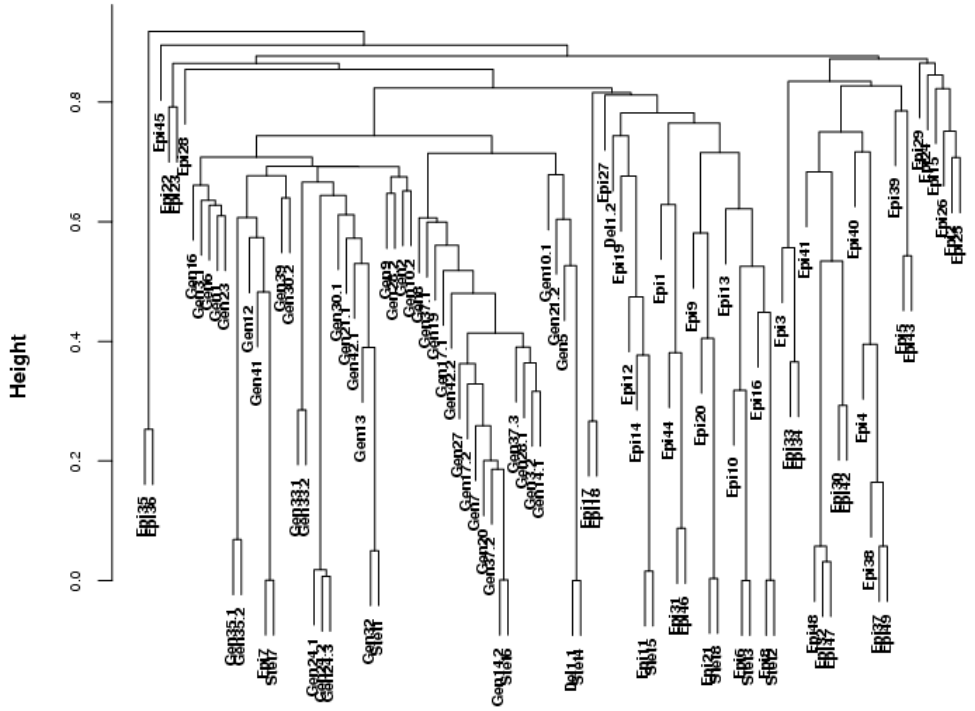

Figure 7. *Dendrogram of S^{block} of the SNP loci and the epidemiological variables of the control group according to i and using S_{τ^2}.*

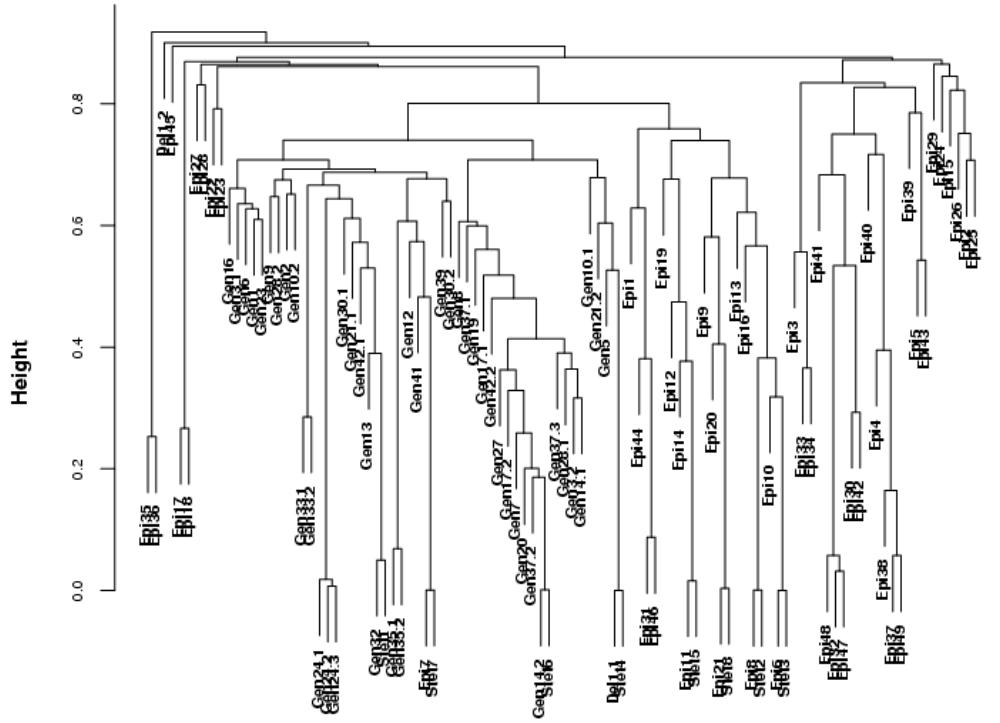

Figure 8. *Dendrogram of S^{block} of the SNP loci and the epidemiological variables of the control group according to i and using S_{τ}.*

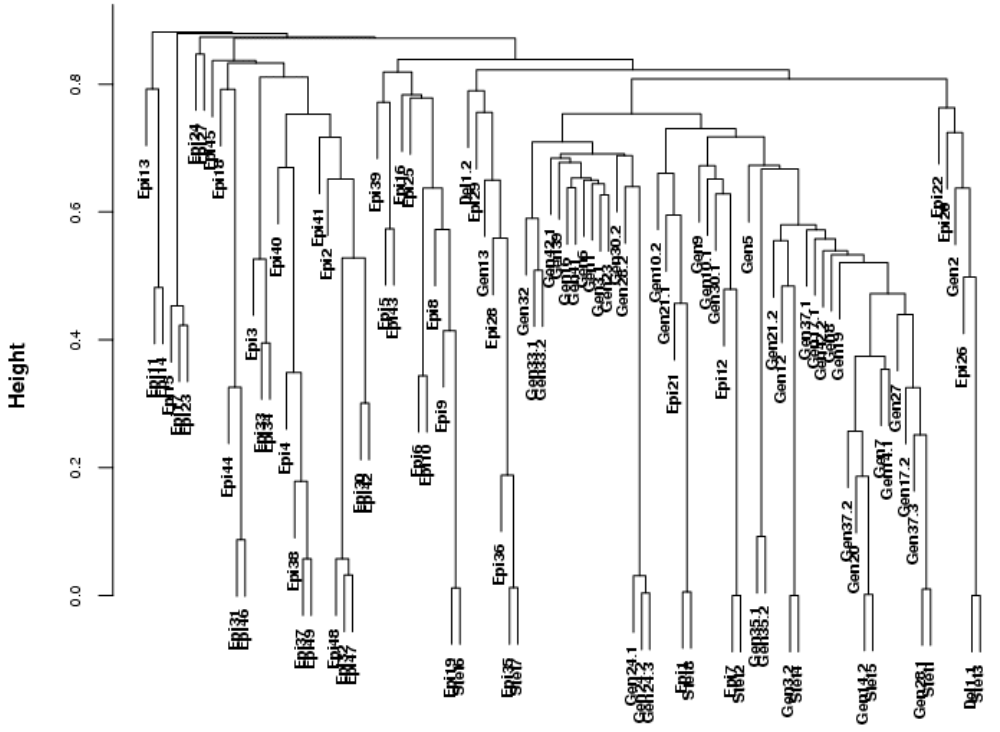

Figure 9. *Dendrogram of S^{block} of the SNP loci and the epidemiological variables of the case group according to i and using S_{τ}.*

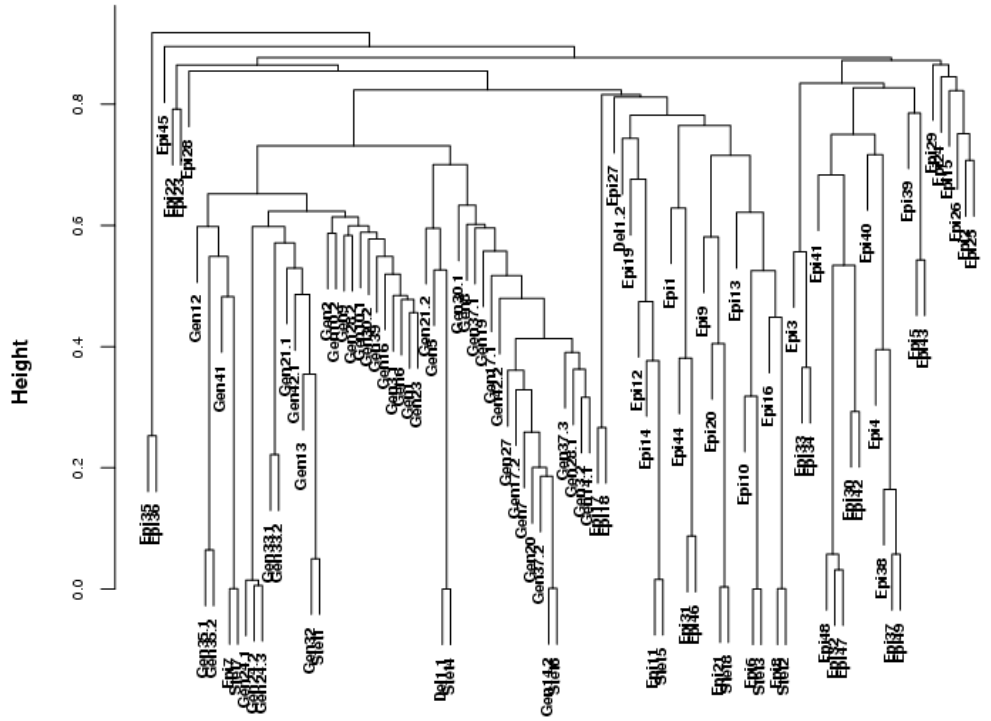

Figure 10. *Dendrogram of S^{block} of the SNP loci and the epidemiological variables of the control group according to ii and using S_{t^2}.*

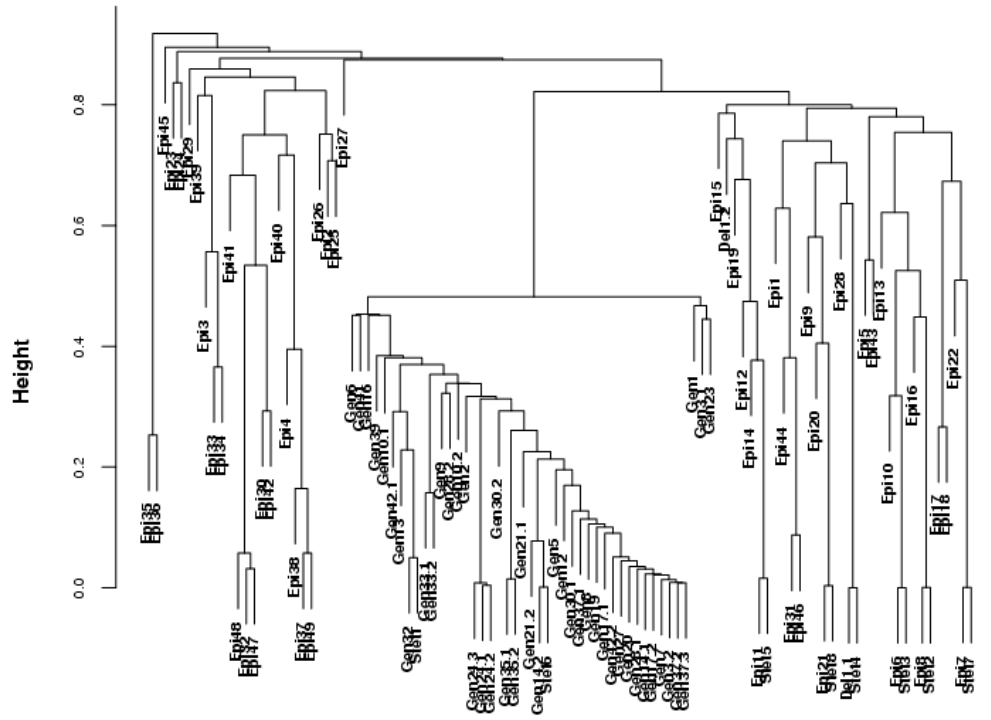

Figure 11. *Dendrogram of S^{block} of the SNP loci and the epidemiological variables of the control group according to iii and using S_{t^2}.*

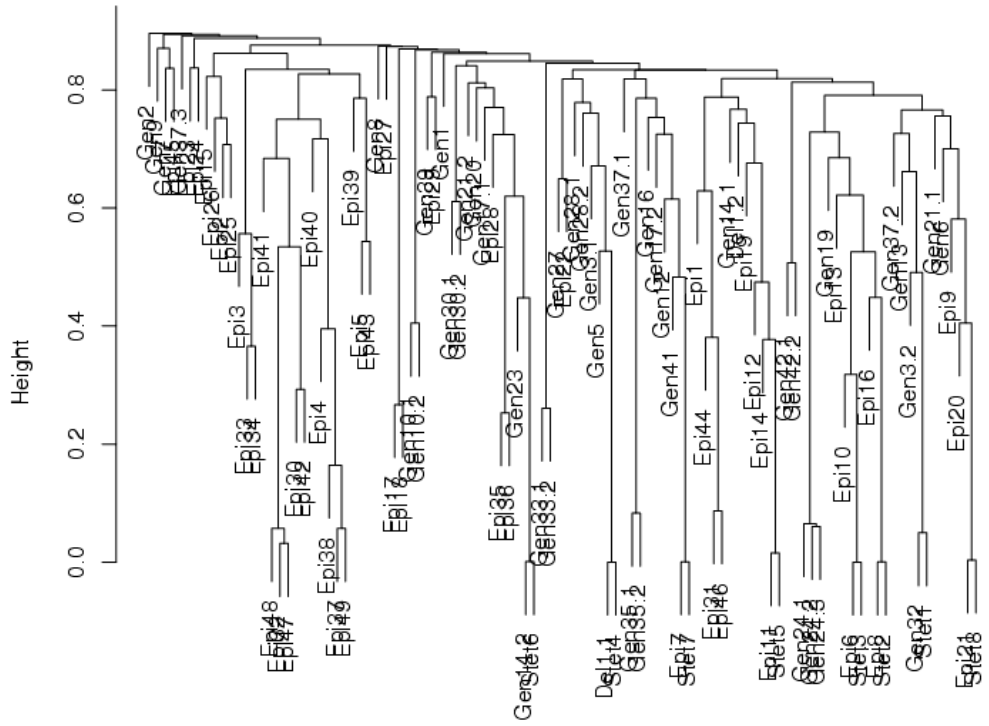

Figure 12. *Dendrogram of S^{block} of the SNP loci and the epidemiological variables of the control group using S_{PC} for all categorial variables and S_{r^2}.*

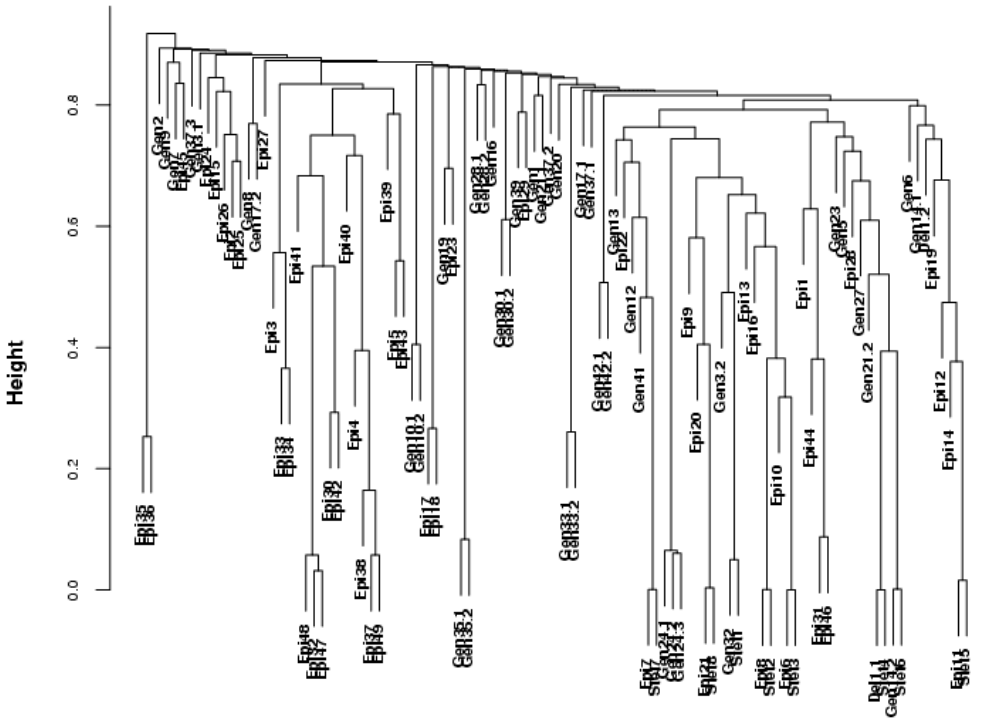

Figure 13. *Dendrogram of S^{block} of the SNP loci and the epidemiological variables of the control group using S_{PC} for all categorial variables and $S_{D'}$.*

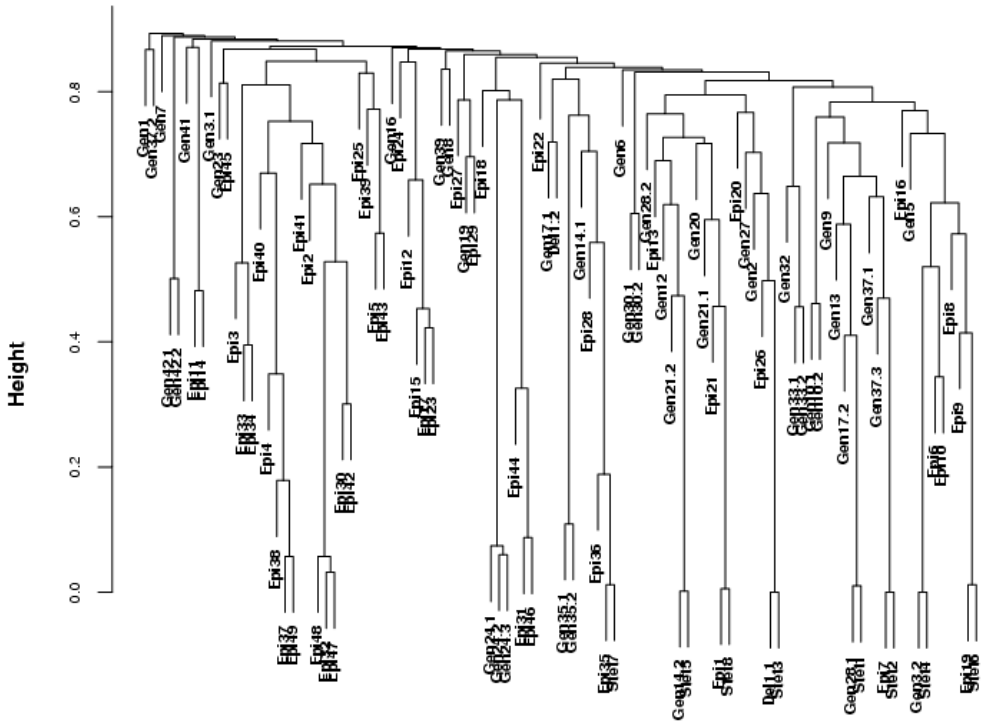

Figure 14. *Dendrogram of S^{block} of the SNP loci and the epidemiological variables of the case group using S_{PC} for all categorial variables and S_{τ^2}.*

In the present case the application of S_{τ^2} and S_{τ} results in almost equal dendrograms (see Figures 7 and 8) with maximal differences applying S_{PC} for all categorial variables (Figures 12 and 13). Note that variables of V^3 are not clustered together but joined together on a high level of similarity to an epidemiologic or genetic variable, Stet1/Gen32 and Stet7/Epi7 in Figure 7, for instance. Generally, the main body of the genetic variables are clustered together in one or two big groups using the Flexible Matching Coefficients for V^1 (Figures 7 – 11). There are no apparent differences between cases and controls concerning small groups of variables, for instance the epidemiologic variables Epi4, Epi37, Epi38, Epi40 and Epi49 as well as Epi32, 47, 48 and Epi3, Epi33 and Epi34 or the genetic variables Gen24.1, Gen24.2 and Gen24.3 (Figures 7 and 9, 12 and 13). Note that loci of the same gene are often clustered together.

Furthermore, the definition of heterozygous and homozygous variants as matches yields mainly the same results here (Figures 7 and 10). The definition of homozygous references and heterozygous loci as matches yields a similar structure with respect to the main body of the epidemiologic variables and with respect to small subgroups (Figures 7 and 11). The genetic variables of V^1 are clustered

together in one poorly structured group (Figure 11). Using S_{PC} for all categorial variables yields a partially different structure though small subgroups of variables remain together, for instance Del1.1, Stet4 and Gen5 (Figures 7 and 12). Generally, differences between cases and controls can mainly be detected in small groups of genetic and epidemiologic variables. Most prominent are the combinations of quantitative and categorial – genetic or epidemiologic – variables. These combinations are consistent within cases and controls, respectively, for all considered combinations of similarity coefficients. For instance, in the control group Epi7 and Stet7 are joined together and form a small cluster with Gen17.2 and Gen41. In the case group Epi7 is clustered together with Stet2 whereas Stet7 is joined together with Epi36 and Epi35 on a similar level of distance. In the control group Stet2 is grouped together with Epi8. In the control group Gen14.2 and Stet6 are clustered together whereas in the case group Stet6 is clustered together with Epi19 and Gen14.2 is clustered together with Stet5. In the control group Stet5 is clustered together with Epi11 which is clustered together with Epi14 in the case group (Figures 7 and 9, Figures 12 and 14).

Clustering of subjects using S^{mixed} with the groups V^g of variables as specified above and $S^1 = S^{flex-IJ,\lambda,\delta}$ as specified above, $S^2 = S^{flex-IJ,\lambda,\delta}$, with $I = \{ii, i = 0, 1, 2, ...\}$, $J = \{ij, i, j = 0, 1, 2, ...\}$, $\lambda_i = 1$, $i = 0, 1, 2, ...$, and $\delta_{ij} = 1$, $i, j = 0, 1, 2, ...$, $S^3 = S_\tau$ and the weights ω^g being the number of variables m_g in subset g yields better structured dendrograms as the use of standard similarity measures for SNP loci only or for SNPs and epidemiologic variables (Figure 15). Due to the large number of subjects interesting clusters can hardly be detected by view, so an automatic search for clusters of a defined minimum size and proportion of cases or controls has to be conducted. Actually, only small groups of cases or control (about 40 – 50) with a slightly elevated proportion of 55 – 60 % cases or controls can be detected.

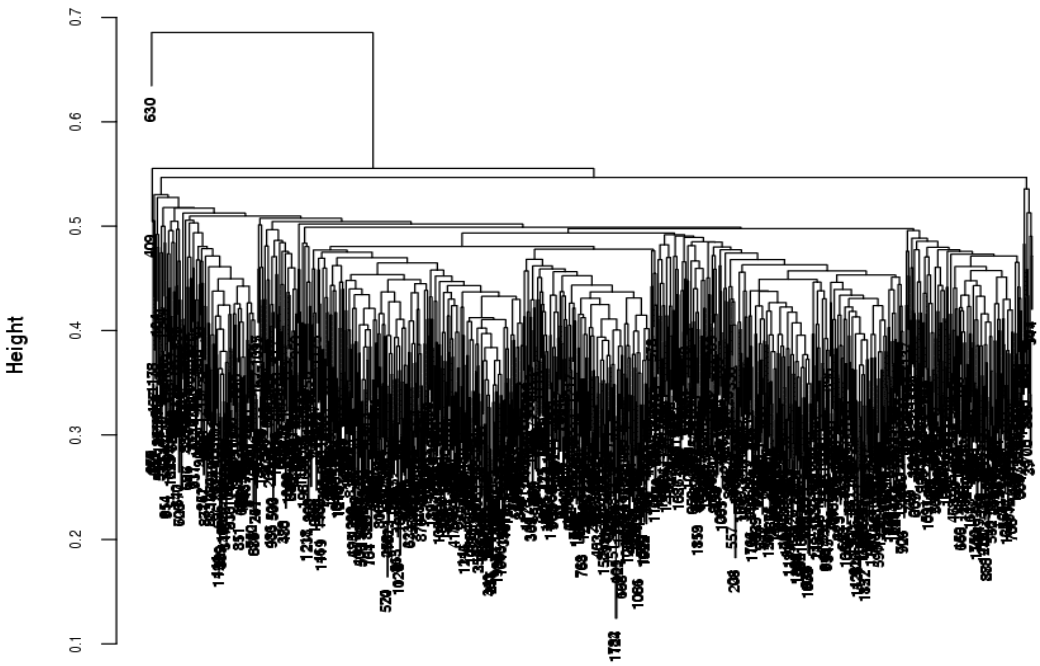

Figure 15. *Dendrogram of S^{mixed} of the SNP loci and the epidemiological variables according to i.*

5. Discussion

The present approach yields interesting hints for potentially relevant combinations of epidemiologic and genetic risk or beneficial factors for sporadic breast cancer. Furthermore, it provides a general insight into relationships between the considered variables that may also be useful for the generation of biological hypotheses. A number of epidemiologic variables can be detected that do not contribute to differences between cases and controls and might be omitted in a further step of the analysis of the data.

Considering the results of the different combinations of similarity coefficients needs further investigation by the use of simulation studies. Especially the impact of standardised or squared correlation coefficients for clustering quantitative variables and the use of the *p*-value of the Kruskal-Wallis test as coefficient of distance of quantitative and categorial variables have to be examined. A further interesting aspect is the impact of the often used categorisation of quantitative variables.

With respect to the clustering of subjects no relevant high or low risk groups have been detected actually. Using Self-Organising Maps several interesting groups with

elevated proportion of cases or controls have been detected (Ittermann, 2006) though it seems to be generally difficult to obtain satisfying results searching for high or low risk groups. Hence, further approaches for clustering subjects have to be examined, for instance the clustering in subspaces (see for instance, Friedman and Meulman, 2004; Parson *et al.*, 2004). The general idea is that different subgroups of objects can be characterised by different subgroups of variables. So, calculating the similarity of two or more objects based on the complete vectors of variables, containing relevant and irrelevant information about their similarity, may hide the "true" structure of the data.

In general, cluster analysis can help to gain more insight into the data but especially in complex data situations it is reasonably combined with further approaches. For the detection of interactions between several gene loci as well as between gene loci and exogenous factors there are a plethora of further approaches. Classification approaches as, for instance, classification trees, ensemble methods, SVM (Schwender *et al*., 2004), multi-dimensionality reduction (MDR) and logic regression (Rabe, 2004) aim to identify those combinations of traits which yield the "best" prediction of the case-control status. The difficulty with these approaches is generally a high misclassification rate due to the heterogeneity of the case-group, the low penetrance of the relevant genetic variants and, hence, the amount of competing models.

So combining cluster and classification approaches – for instance, by a pre-selection of variables or by joint hints towards of potential impact factors by several approaches – may help to gain more insight and to develop new biological hypothesis.

Acknowledgements

The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") is gratefully acknowledged.

The authors thank all partners within the GENICA (Interdisciplinary Study Group on Gene Environment Interaction and Breast Cancer in Germany) research network (represented by C. Justenhoven, Stuttgart, H. Brauch, Stuttgart, S. Rabstein, Bochum, B. Pesch, Bochum, V. Harth, Bonn/Bochum, U. Hamann, Heidelberg, T. Brüning, Bochum, Y. Ko, Bonn) for their cooperation.

References

Anderberg MR (1973). *Cluster analysis for applications*. Academic Press, New York.

Beral, V (2003). Breast cancer and hormone-replacement therapy in the Million Women Study. *The Lancet* **362**, pp. 419-427.

Cox TF, Cox MAA (2001). *Multidimensional Scaling*, 2nd ed. Chapman & Hall /CRC, Boca Raton, Florida, USA.

Eisen MB, Spellman PT, Brown PO, Botstein D (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, pp. 14863-14868.

Friedman JH, Meulman JJ (2004). Clustering Objects on Subsets of Attributes. *Journal of the Royal Statistical Society, Series B* **66**, pp. 1-25.

Garte S (2001). Metabolic susceptibility genes as cancer risk factors: Time for a reassessment? *Cancer Epidemiology, Biomarkers & Prevention* **10**, pp. 1233-1237.

Hartung J, Elpelt B, Klösner K-H (1991). *Statistik*. 8th ed. R. Oldenbourg Verlag, München.

Hastie T, Tibshirani R, Botstein D, Brown P (2001). Supervised harvesting of expression trees. *Genome Biology* **2**, pp. 1-12.

Hollander M, Wolfe DA (1999). *Nonparametric statistical methods.* Wiley, New York, USA.

Ittermann T (2006). *Analyse von SNP und epidemiologischen Daten mittels Self-Organizing Maps.* Diploma thesis, University of Dortmund.

Kornrumpf J (1986). *Hierarchische Klassifikation einer Objektmenge.* Peter Lang, Frankfurt a.M.

Müller T (2004). *Clusteranalyse von SNP Daten: Verschiedene Ähnlichkeitsmaße im Vergleich*. Diploma thesis, University of Dortmund.

Müller T, Selinski S, Ickstadt K (2005). Cluster analysis: A comparison of different similarity measures for SNP data. *Technical Report 14/05*, University of Dortmund.

Ostermann R, Degens PO (1984). Eigenschaften des Average-Linkage-Verfahrens anhand einer Monte-Carlo-Studie. *In:* H.-H. Bock (Ed.): *Anwendungen der Klassifikation: Datenanalyse und numerische Klassifikation.* Indeks Verlag, Frankfurt, pp. 108-114.

Parson L, Haque E, Liu H (2004). Subspace Clustering for High Dimensional Data: A Review. *SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining* **6**, pp. 90-105.

Rabe C (2004). *Identifying interactions in high dimensional SNP data using MDR and Logic Regression*. Diploma Thesis, University of Dortmund.

Schwender H, Zucknick M, Ickstadt K, Bolt HM (2004). A Pilot Study on the Application of Statistical Classification Procedures to Molecular Epidemiological Data. *Toxicology Letters* **151**, pp. 291-299.

Selinski S, Ickstadt K (2005). Similarity measures for clustering SNP data. *Technical Report 27/05*, University of Dortmund.

Sitterberg G (1978). Zur Anwendung hierarchischer Klassifikationsverfahren. *Statistische Hefte* **19**, pp. 231-246.

Steinhausen D, Langer K (1977). *Clusteranalyse*. Walter de Gruyter, Berlin.

Zhang B, Srihari SN (2002). A fast algorithm for finding *k*-Nearest Neighbors with non-metric dissimilarity. *Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR'02)*.