Similarity Measures for Clustering SNP Data

Silvia Selinski and Katja Ickstadt

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 particular kind of genetic data – so called SNP data – 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 65 SNP loci and 2 insertions of longer sequences in genes involved in the metabolism of hormones, xenobiotics and drugs as well as in the repair of DNA and signal transduction. 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 – we aim 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. We consider here the problem of suitable measures of proximity of two variables or subjects as an indispensable basis for a further cluster analysis.
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. Comparing the individual genetic profiles as well as comparing the genetic information across subpopulations we discuss possible choices of similarity measures, in particular similarity measures based on the counts of matches and mismatches. New matching coefficients are introduced with a more flexible weighting scheme to account for the general problem of the comparison of SNP data: The large proportion of homozygous reference sequences relative to the homo- and heterozygous SNPs is masking the accordances and differences of interest.

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

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.

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.

The search for patterns in the present data set may be performed by a variety of clustering and classification approaches. We 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 will be considered in particular due to their flexibility and their suitability for the present problem. Besides the usual matching coefficients new ones are introduced that may account for biological background knowledge or hypothesis due to their flexible weighting scheme. Furthermore they are able to handle a special feature of SNP data: The proportion of homo- or heterozygous SNPs is usually rather small compared to the proportion of homozygous reference sequences, i.e. loci that contain no sequence variation. So in comparing two variables or subjects there is a huge amount of common homozygous reference sequences, which we denote as *0-0-matches*, masking the interesting differences or similarities: the small amount of common or mismatching homo- and heterozygous polymorphisms.

The most common genetic data are actually microarray data measuring gene expression levels of thousands of genes simultaneously and there are numerous publications dedicated to the issue of clustering this type of data

on the basis of measures of proximity, e.g. Brazma & Vilo (2000), Eisen *et al.* (1998), Hastie *et al.* (2001), Tibshirani *et al.* (2001). Roughly speaking, gene expression levels give a measure of the activity of the considered genes on a continuous scale, in contrast to SNPs data, where the information about the inherited variants of these genes is considered. Thus, for gene expression data measures of proximity for qualitative data based on the concept of correlation as well as metrics, the Euclidean distance, for instance, can be used.

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 more that 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.

Hence, the question is, how to assign a numerical value measuring the proximity – similarity or dissimilarity – of two SNP loci or of the genetic profile of two persons?

There are plenty of potential similarity or distance measures for this attempt (see e.g. Cox & Cox, 2001). After an introduction to the biological background of the GENICA study we will give an overview over possible approaches. The conventional matching coefficients are extended to a new class of more flexible matching coefficients. Chapter 5 gives some of the results for these Flexible Matching Coefficients. A detailed comparison of the introduced coefficients of similarity is given by Müller *et al.* (2005).

2. Background

The problem of measuring the proximity of genetic data arises in many studies as for example in the GENICA study of sporadic breast cancer. GENICA is part of the German Human Genome Project (DHGP) and is dedicated to the investigation of genetic interactions and gene-environment interactions leading to sporadic breast cancer.

2.1 Sporadic breast cancer

In Germany almost 50 000 women develop breast cancer each year, that are 7 to 10 % of the female population developing this disease during their life-time. Breast cancer is the most frequent cancerous disease in women with about 26 % of all newly detected cancers. About one third of all patients are younger than 60 years while a tendency towards a more frequent development of breast cancer is reported generally and especially for younger women (ZTG, 2004).

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. Several exogenous risk factors seem to influence the risk of sporadic breast cancer. 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).

This is important not only for future developments of therapeutic approaches but also for prevention and earlier diagnosis and, hence, for a better prognosis. Thus, identification of high risk combinations of genetic and

exogenous factors would facilitate prevention and permit intensification of medical check-ups for women with high risk profiles.

For the genetic basis of cancer in general and sporadic breast cancer in particular, see, for instance, Snustad & Simmons (1999) and Rabe (2004).

2.2 Genetic terms

The genetic information of all living organisms, except some viruses, is stored in DNA (deoxyribonucleic acid). Generally, nucleic acids are macromolecules composed of repeating subunits, the so called nucleotides. Each nucleotide is composed of a phosphate group, a five-carbon sugar or pentose and a cyclic nitrogen-containing base. In DNA (deoxyribonucleic acid), the sugar is 2-deoxyribose and in RNA (ribonucleic acid), the sugar is ribose. The four bases in DNA are: adenine (A), guanine (G), cytosine (C), and thymine (T). The bases in RNA are the same except that RNA contains uracil (U) instead of thymine (T). Adenine and guanine are double-ring bases called purines. Cytosine, thymine and uracil are single-ring bases called pyrimidines. Thus, DNA and also RNA are composed of four different nucleotides, two purines and two pyrimidines, which are joined together in long chains. RNA is usually found as a single stranded polymer whereas DNA is organised as a double-stranded helix. The two strands of a DNA double helix are said to be complementary because of the specific base-pairing: adenine is always paired with thymine, and guanine is always paired with cytosine (see Figure 1). Thus, all base pairs consist of one purine and one pyrimidine. The DNA macromolecules are organised in chromosomes. In humans the diploid set of chromosomes is 46: two

homologous sets – one maternal and one paternal – of 22 autosomes and one sex chromosome.

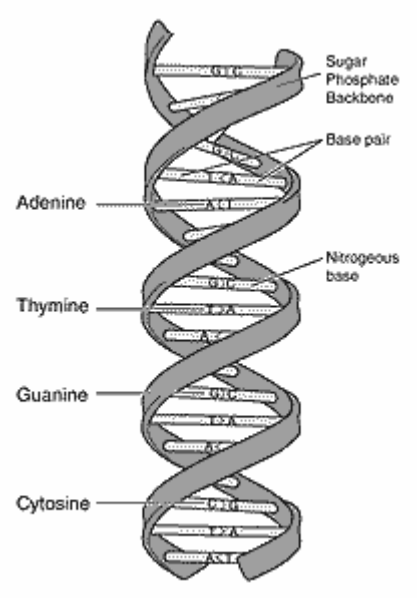

Figure 1. *DNA double helix.*

The expression of the genetic information involves mainly two steps: transcription and translation (see Figure 2). First, one strand of the DNA is used as a template to synthesize a complementary strand of RNA: the gene transcript. This process is called transcription and occurs in the nucleus of the cell. Transcription is initialised at specific nucleotides sequences called promoters which are located before the transcription start point. The efficiency of a promoter is influenced by nearby enhancer sequences.

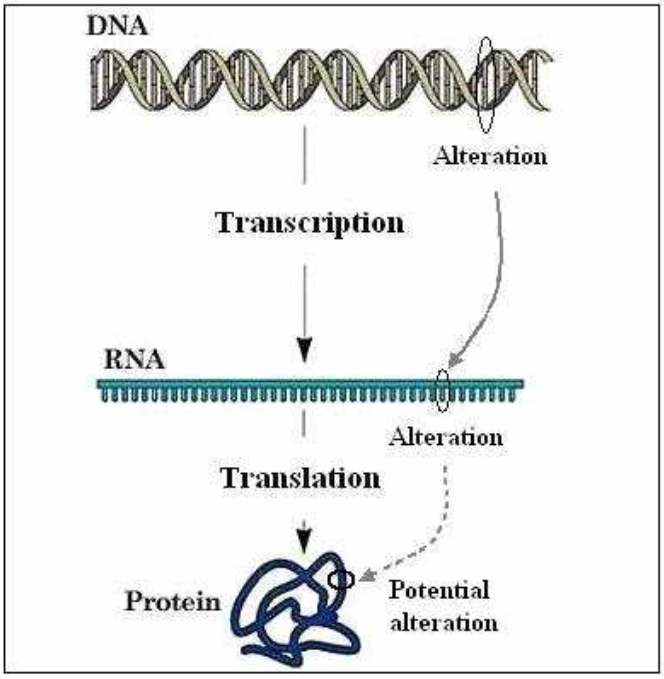

Figure 2. *Transmission of the genetic information from DNA to protein.*

Most genes that code for proteins are so called split genes. That means they contain coding sequences – exons – and non-coding sequences – introns. The biological significance of the latter remains unclear. Each intron must be removed from the RNA transcript of a gene before translation. This process is called splicing and has to be very precise to assure that codons in exons may be read correctly during translation. Multiple introns of a gene can be removed separately or in combination depending on how the splicing machinery interacts with the RNA. Joint excision of two introns means that also the exon in between will be removed. Thus, the coding sequence of an RNA can be modified by deleting some of its exons. This phenomenon of splicing an RNA transcript in different ways, called alternate splicing, makes it possible for a gene to encode different polypeptides (Snustad and Simmons, 1999).

After RNA transcript processing the so called mRNA (messenger RNA) is transferred to the cytoplasm. During translation the sequence of nucleotides in the RNA transcript is converted into the sequence of amino acids in the polypeptide gene product.

This conversion is conducted by the genetic code: the specification of the 20 amino acids by nucleotides triplets called codons. Each but three of the 64 triplets codes for a specific amino acid, the three further are polypeptide chain termination – or stop – codons (see Table 1). Most amino acids are specified by more than one codon, with similar amino acids being specified by related codons. The first and the second nucleotide of a codon are the most important 'letters' for amino acid specification as many base substitutions at the third position do not change the specified amino acid. Moreover, amino acids with similar chemical properties have codons that differ from each other by only one base. Thus, many single base pair substitutions will result in gene products that minimize the effect of mutations (see Tables 1 and 2).

Table 1. *The genetic code according to Snustad and Simmons (1999). 'Stop' denotes a terminator. Abbreviations are given in table 2.*

1st (5') letter		2nd letter: U		2nd letter: C		2nd letter: A		2nd letter: G		3rd (3') letter
U	U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
		UUC		UCC		UAC		UGC		C
		UUA	Leu	UCA		UAA	Ochre (stop)	UGA	Opal (stop)	A
		UUG		UCG		UAG	Amber (stop)	UGG	Trp	G
	C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
		CUC		CCC		CAC		CGC		C
		CUA		CCA		CAA	Gln	CGA		A
		CUG		CCG		CAG		CGG		G
	A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
		AUC		ACC		AAC		AGC		C
		AUA		ACA		AAA	Lys	AGA	Arg	A
		AUG	Met (initiator)	ACG		AAG		AGG		G
	G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
		GUC		GCC		GAC		GGC		C
		GUA		GCA		GAA	Glu	GGA		A
		GUG		GCG		GAG		GGG		G

Table 2. *Abreviations and groups of amino acids according to Snustad and Simmons (1999).*

Hydophobic or nonpolar side groups	
Gly	Glycine
Ala	L-Alanine
Val	L-Valine
Leu	L-Leucine
Ile	L-Isoleucine
Pro	L-Proline
Phe	L-Phenylalanin
Met	L-Methionine
Trp	L-Tryptophan
Cys	L-Cystein
Acidic side groups	
Asp	L-Aspartic acid
Glu	L-Glutamic acid

Hydrophilic or polar side groups	
Ser	L-Serine
Thr	L-Threonine
Tyr	L-Tyrosine
Asn	L-Asparagine
Gln	L-Glutamine
Basic side groups	
Lys	L-Lysine
Arg	L-Arginine
His	L-Histidine

The expression of genes is regulated via regulation of the transcription of genes, via processing regulation that involves alternate splicing or via regulation of the translation involving mRNA stability.

The human genome consist of about 3 billion base pairs and about 30 000 genes. We share about 99.9% of our DNA. Thus, about 3 million sequence differences can be detected comparing two individuals. Genetic variations include mutations and polymorphisms. A polymorphism is a genetic variation that is present in at least 1% of a population. The most common form of genetic variation – about 90% – are so called single nucleotide polymorphisms (SNPs) that are expected to occur every 1000 base pairs.

To be precise, a SNP (single nucleotide polymorphism) is a point mutation that is present in more that 1 % of a population. A point mutation is a change of one base pair with respect to the most frequent variant, or a deletion that means the respective base pair is missing, or an addition of one base pair (see Figure 3).

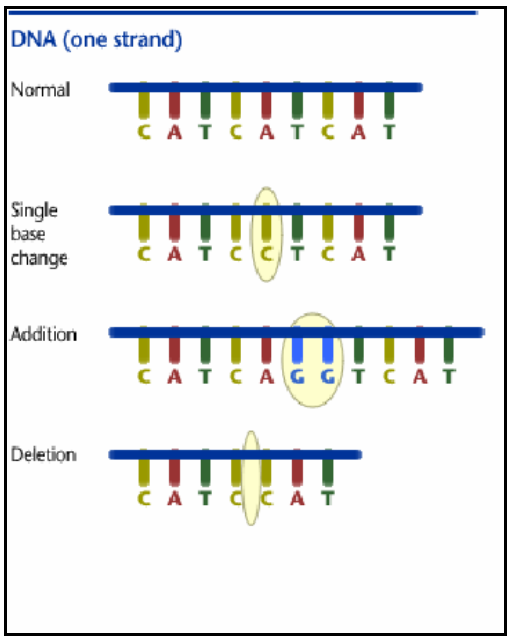

Figure 3. *Possible point mutations.*

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. The most frequent variant is also called *major allele* or *reference sequence*, the less frequent *minor allele* or *variant*. The most frequent point mutation is the transition, the substitution of one purine (A, G) base by the other one or the substitution of one pyrimidine base (C, T) by the other one, respectively. The transversion, that means the substitution of a purine base by a pyrimidine base or vice versa, as well as deletions or additions occur less frequently.

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 (*non-synonymous exchange*) or the switch from an amino acid coding triplet to a stop codon leading to a shortened amino acid chain (Figure 2). The impact of such an alteration of the amino acid chain depends on its position, for example if an exchange of one amino acid occurs in a functional region of an enzyme. Though some of the SNPs do not result in an amino acid exchange (*synonymous exchange*) an effect is not always deniable. With respect to SNPs that are located in non-coding regions single alterations of the sequence may have an impact on gene regulation, for instance.

Though most of these polymorphisms are supposed to have generally a minor impact, under certain environmental conditions some have indeed an effect contributing, for instance, to the development of a disease. A

prominent example is the genetic variation of N-acetyltransferase-2 (NAT2) where single nucleotide polymorphisms of the gene result in phenotypically slow acetylator types which in turn are more susceptible to environmental and industrial carcinogens. For instance, slow acetylators are at higher risk of developing bladder cancer due to occupational exposure to aromatic amines than fast acetylators as the detoxification of these substances is less effective (Thier *et al.*, 2003).

3. Data

The present data set consists of a selection of SNP loci of the GENICA study of sporadic breast cancer. The GENICA study is a population-based age-matched case-control study assessing genotypes of over 120 SNP loci and exogenous risk factors of the reproductive history, hormone use, life style factors, occupational history, family history of cancer, etc. of 1100 cases and 1100 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).

Actually the available data set comprises 65 SNP loci and 2 loci where the variant sequence is an insertion of 306 and 16 base pairs, respectively, of 610 cases of sporadic breast cancer and of 650 age-matched healthy controls from the first phase of recruitment.

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. Pathway means the field where a gene-product plays a role within the human metabolism, e.g. the pathway of xenobiotics and drug metabolism. Table 3 gives the considered pathways and the corresponding genes. Note that a gene may participate in more than one pathway.

Table 3. *Assignment of the considered genes to their pathways and number of investigated loci per pathway*

Pathway	Gene	Number of SNPs
Metabolism of xenobiotics and drugs	12, 14, 17, 18, 19, 23, 27, 79, CYP1A1, CYP1B1, CYP2E1, GST, NAT, ADH2	30
Metabolism of steriod hormones	23, 34, 53, 58, 100, 101, 102, 105, CYP1A1, CYP1B1	12
DNA repair	24, 25, 55, 72, 74	7
Nutrition relevant factors	32, 45, 62	6
Signal transduction	33, 64, 75, 76, 77, 78, 80, 81	9
Growth factors	70	2
Oncogene	31	2
Transporter	38	3
Detoxification	41	1
Control of cell cycle	103, 104	4

Part of the locus names are coded due to their origin from different institutes.

The data set comprises 47 transitions – 27 exchanges of guanine and adenine, 20 exchanges of cytosine and thymine, 22 transversions – the most frequent exchange was between cytosine and guanine with 11 loci, 3 deletions and 2 insertions.

4. Methods

Searching for patterns in the data there are two possible objectives: a comparison of variables or a comparison of subjects. In the first case we aim 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 with respect to their genetic information with the aim of finding high and low risk groups. 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.

There are lots of measures of proximity representing different concepts of similarity and different assumptions with respect to the data. Introducing first the general concepts of proximity for the particular situation of SNP data we then give a short summary of the definition and properties of similarity measures. 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 genetic alterations. In section 4.3 we present different classes of measures of similarity and discuss their appropriateness for the present data structure. A new family of matching coefficients – the Flexible Matching Coefficients - which accounts for the special features of SNP data set and biological assumptions is introduced in section 4.3.1.

4.1 Concepts of proximity

Focusing on the similarity of the genetic variables the basic question is: What does similarity of two SNP loci mean and how to measure it?

The present data base contains genotypes and some additional information about the SNP loci. So, basis of a search for patterns is a comparison of genotypes of different loci in the same or different genes.

One possibility is to consider the similarity of loci with respect to their joint occurrence of similar genotypes. In a first step we here need to assess the number of persons carrying the respective combination for each combination of genotypes, e.g. the number of persons who are heterozygous at locus A and show the homozygous reference sequence at locus B (see Figure 4).

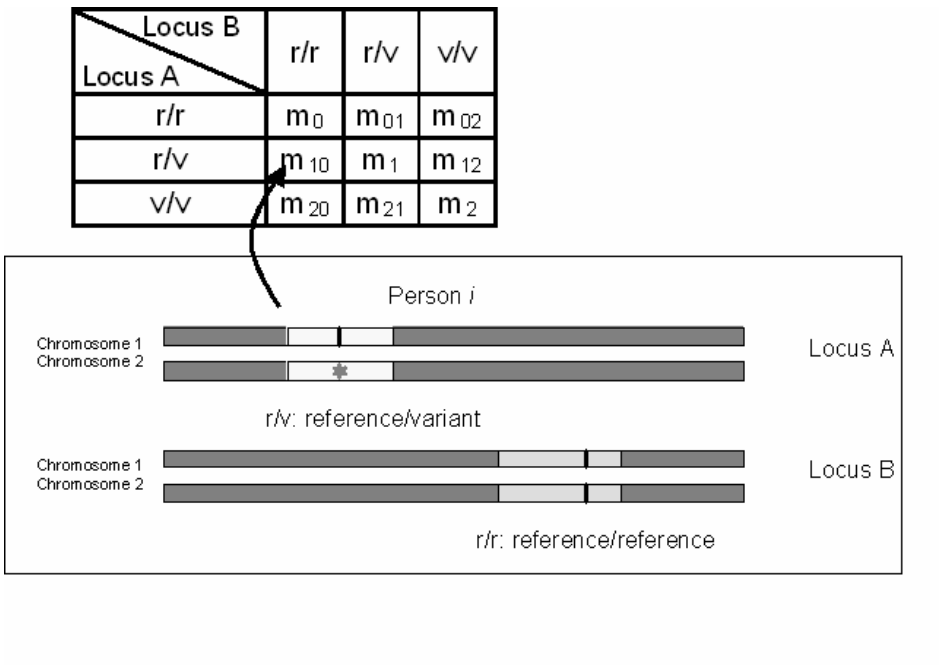

Figure 4. *Comparison of two SNP loci. The SNP loci are indicated by a black bar in case of the reference base pair and by a grey star in case of the variant.*

The second step is to determine which genotypes of the two loci we regard as similar. A similar combination is obviously the joint occurrence of homozygous reference types. For all other combinations of genotypes at two loci it is not that obvious which ones are similar and raises the question of the consequences of a homo- or heterozygous SNP at a particular locus compared with the reference sequence.

Thus the general idea is to consider the potential deviations of the gene products from their most frequent variant to which each of the investigated loci contribute.

Comparing the genotypes of two SNP loci with respect to their impact on the metabolism would mean that we have to assess first a numerical value to each hetero- and homozygous variant characterising their effect compared to the homozygous reference variant, e.g. their contribution to risks, beneficial effects, influence on gene-regulation, ensuring that these numerical values may be compared across all loci. Hence, concepts of similarity based on correlation or deviation from independence would be the appropriate approach to search for patterns of SNP loci.

Thus considering the rather 'rough' standardised information about the hetero- or homozygous deviation from the reference sequence, interpreting these data as information about the amount of 'original gene-dose' and thus drawing conclusions about the potential impact is a reasonable approach. Anyway, considering specific measures of similarity it is possible to incorporate further biological assumptions, potential benefits of heterozygosity, for instance, or the existence of at least one reference copy of a gene that may code for the 'most common' enzyme variant. Similarity may be considered then in terms of *agreement* or in terms of *dependence*.

Agreement means to consider two loci as similar if the majority of subjects owns a combination of similar genotypes at these loci. Two loci would be considered as dissimilar if the majority of subjects has a combination of dissimilar genotypes. Matching coefficients and measures of correlation, for instance, would correspond to this concept of similarity.

The concept of dependence encompasses the first in so far as a frequent occurrence of similar genotypes would also be regarded as similarity. But it also allows generally for further combinations of genotypes – perhaps a priori judged as dissimilar – to contribute to the label ‘similar’ for two SNP loci 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 and measures based on the χ^2-statistic.

Focussing on the comparison of objects or subjects (observed persons in our example) means to assess the similarity of the individual genotypes at each locus and to draw conclusions about the overall similarity of all considered loci. Generally, two subjects can be considered as similar if they share similar genotypes at most loci. They are dissimilar if most considered loci show dissimilar combinations of genotypes (see Figure 5). This raises again the question of similarity of the observed genotypes but here the similarity of genotypes at a single locus.

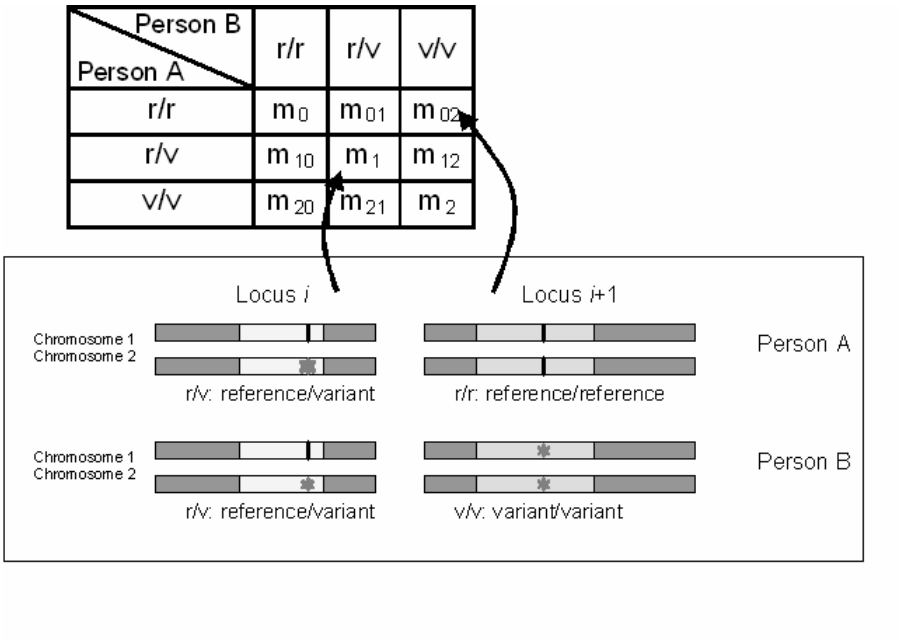

Figure 5. *Comparison of two persons at two loci. The SNP loci are indicated by a black bar in case of the reference base pair and by a grey star in case of the variant.*

Considering first the similarity of genotypes at a particular gene locus implies to consider the consequences of sequence alterations – homo- and heterozygous – with respect to their reference, for instance the loss of function of an enzyme in the drug metabolism. Unless the potential consequences of single alterations encompass a broad range of effects – as stated above – we have to concentrate primarily on the information about the sequence variants. In a further step it is possible to incorporate knowledge about inheritance, basic information about the relevance of homo- and heterozygosity of the alterations and assumptions about the relevance of loci and genes using different weighting schemes.

Generally, two persons can be considered as similar if they share the same genotype at most loci. Thus similarity means here *accordance* or *agreement*.

The concept of *dependence* is less adequate. Imagine that 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.

4.2 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 $S : V \times V \rightarrow IR$, with V being the set of variables by $S : O \times O \rightarrow IR$, 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 $V = \{V_1, \ldots, V_m\}$. Then a measure of similarity of two variables $V_k \in V$ and $V_l \in V$, is given by $S : V \times V \rightarrow IR$ 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 $D : V \times V \rightarrow IR$ 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 $D : V \times V \to [0,1]$.

4.3 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 for each concept of similarity: *agreement* and *dependence*, as well as measure for quantitative data based on a geometric interpretation of proximity (*distance*). We relate the different situations to the present problem introducing new measures developed for this particular data situation.

4.3.1 Nominal scale

A special case of nominal scaled data is binary data, for instance the presence or absence of a trait. As many measures for categorial data are derived from the binary case and the transformation of data to a binary scale is a common approach we introduce first measures of agreement for this particular kind of data. We continue with the general case of $p \geq 2$ categories and extend the usual matching coefficients to a more general family of matching coefficients: Flexible Matching Coefficients. Measures of dependence which are able to cope with different numbers of categories are introduced generally for both cases: binary data and $p \geq 2$ categories.

Measures of agreement – Special case: Binary data

Considering the present data situation the information about the SNP loci might be transformed to a binary scale by introducing for each locus two new variables denoting

i. the occurrence of at least one SNP and

ii. the occurrence of at least one reference sequence.

Thus, a homozygous reference would result in ‘0’ for the first variable and ‘1’ for the second variable and vice versa in case of a homozygous SNP. Heterozygosity would then be denoted by ‘1’ for both variables. Assuming that one of these two variables is rather less informative, for instance if homo- and heterozygous references are considered as quite similar, one may omit one of these two variables reducing the information to two categories. A binary representation may be used for both: a comparison of variables and a comparison of subjects.

A special problem of the present data situation is the huge amount of homozygous reference types. Thus comparing two variables or subjects the proportion of combinations of homozygous references, further called *0-0-matches*, is rather high compared with the remaining combinations and might be supposed to mask the interesting effects. The term 0-0-matches arises from matching coefficients for binary data, where the number of common presence – *1-1-matches* – and common absence – 0-0-matches – of a trait is related to the number of mismatching combinations, i.e. one absence and one presence of a trait.

Denote V_k and V_l being two variables that should be compared with respect to their similarity and m_{00}, m_{01}, m_{10}, m_{11} as given by Table 4. The case of two objects that should be compared substitute V_k and V_l by O_k and O_l, respectively.

Table 4. *Contingency table of V_k and V_l, m_i and m_{ij} denoting the respective numbers of combinations of categories i and j.*

V_k \ V_l	0	1
0	m_{00}	m_{01}
1	m_{10}	m_{11}

Hence, all of the following measures can be derived from the corresponding table of contingency.

Most measures of agreement can be generalized to (Steinhausen & Langer, 1977)

$$S^{\lambda,\delta} = \frac{m_{11} + \lambda m_{00}}{m_{11} + \lambda m_{00} + \delta(m_{10} + m_{01})} \qquad (1)$$

where λ = 1, if the measures does not make any difference between 0-0- and 1-1-matches and λ = 0, if the measure treats the 0-0-matches as an uninformative absence of a trait not contributing to the similarity or dissimilarity of two variables or objects. Furthermore matches and mismatches are weighted differently depending on the value of $\delta > 0$.

An overview over common matching coefficients is given in Table 5, see also Anderberg (1973) and Cox and Cox (2001) for details and for further matching coefficients.

Table 5. *Matching coefficients for binary data.*

Symbol	Coefficient	Name
Measures of Similarity including the 0-0-matches		
$S_M^{1,1}$	$\frac{m_{11}+m_{00}}{m_{11}+m_{00}+m_{10}+m_{01}}$	Simple Matching
$S_{SoSn}^{1,1/2}$	$\frac{m_{11}+m_{00}}{m_{11}+m_{00}+\frac{1}{2}(m_{10}+m_{01})}$	Sokal & Sneath
$S_{RT1}^{1,2}$	$\frac{m_{11}+m_{00}}{m_{11}+m_{00}+2\cdot(m_{10}+m_{01})}$	Rogers & Tanimoto I
S_{K1}	$\frac{m_{11}+m_{00}}{m_{10}+m_{01}}$	Kulczynski I
S_{H1}	$\frac{m_{11}+m_{00}-(m_{10}+m_{01})}{m_{11}+m_{00}+m_{10}+m_{01}}$	Hamman I
S_{Phi}	$\frac{m_{11}\cdot m_{00}-m_{10}\cdot m_{01}}{[(m_{11}+m_{10})\cdot(m_{11}+m_{01})\cdot(m_{00}+m_{10})\cdot(m_{00}+m_{01})]^{1/2}}$	Phi
S_Q	$\frac{m_{11}\cdot m_{00}-m_{10}\cdot m_{01}}{m_{11}\cdot m_{00}+m_{10}\cdot m_{01}}$	Yule *Q*
S_Y	$\frac{\sqrt{m_{11}\cdot m_{00}}-\sqrt{m_{10}\cdot m_{01}}}{\sqrt{m_{11}\cdot m_{00}}+\sqrt{m_{10}\cdot m_{01}}}$	Yule *Y*
	Measures of Similarity excluding the 0-0-matches	
$S_J^{0,1}$	$\frac{m_{11}}{m_{11}+m_{10}+m_{01}}$	Jaccard
$S_D^{0,1/2}$	$\frac{m_{11}}{m_{11}+\frac{1}{2}(m_{10}+m_{01})}$	Dice
$S_{RT2}^{0,2}$	$\frac{m_{11}}{m_{11}+2\cdot(m_{10}+m_{01})}$	Rogers & Tanimoto II
S_{K2}	$\frac{m_{11}}{m_{10}+m_{01}}$	Kulczynski II
S_{H2}	$\frac{m_{11}-(m_{10}+m_{01})}{m_{11}+m_{10}+m_{01}}$	Hamman II
S_O	$\frac{m_{11}}{[(m_{11}+m_{10})\cdot(m_{11}+m_{01})]^{1/2}}$	Ochiai

Measures that can be derived from eq. (1) are restricted to [0,1] (see Remark 5). The measures of Hamman, Phi and both coefficients of Yule

may result in negative values. Except the coefficients of Kulczynski all measures of similarity shown in table 5 do not exceed 1.

Measures of agreement for categorically scaled data

Consider the general 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.

Extensions are possible and considered in detail in the next section.

Hence, most measures for binary data can be extended to more than two categories without any problems. The special role of the 0-0-matches persists extending the binary case to the p categorical case (see also Steinhausen & Langer, 1977).

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 6.

Table 6. *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 facilitation, let

$$m^+ = \sum_{i=0}^{p} m_{ii} \quad \text{be the number of matches and} \tag{2}$$

$$m^- = \sum_{i=0}^{p} \sum_{\substack{j=0 \\ i \neq j}}^{p} m_{ij} \text{ be the number of mismatches.} \tag{3}$$

Most measures of agreement have the general form

$$S^{\lambda,\delta} = \frac{m^+ - (1-\lambda)m_0}{m^+ - (1-\lambda)m_0 + \delta m^-}, \tag{4}$$

with $\lambda = 1$ if the 0-0-matches are treated as normal matches, $\lambda = 0$ if the 0-0-matches denote the common absence of a trait and are excluded from the calculation of the similarity between two variables or objects and $\delta > 0$ denoting the weight of the mismatches.

In the special case of SNP data with 3 categories this is

$$S^{\lambda,\delta} = \frac{m_2 + m_1 + \lambda m_0}{m_2 + m_1 + \lambda m_0 + \delta(m_{21} + m_{12} + m_{20} + m_{02} + m_{10} + m_{01})}. \tag{5}$$

An overview over the most common measures is given in Table 7.

Table 7. *Matching Coefficients for categorial data with p ≥ 2 categories*

Symbol	Coefficient	Name
Measures of Similarity including the 0-0-matches		
$S_M^{1,1}$	$\frac{m^+}{m^+ + m^-}$	Simple Matching
$S_{SoSn}^{1,1/2}$	$\frac{m^+}{m^+ + \frac{1}{2}m^-}$	Sokal & Sneath
$S_{RT1}^{1,2}$	$\frac{m^+}{m^+ + 2 \cdot m^-}$	Rogers & Tanimoto I
S_{K1}	$\frac{m^+}{m^-}$	Kulczynski I
S_{H1}	$\frac{m^+ - m^-}{m^+ + m^-}$	Hamman I
	Measures of Similarity excluding the 0-0-matches	
$S_J^{0,1}$	$\frac{m^+ - m_0}{m^+ - m_0 + m^-}$	Jaccard
$S_D^{0,1/2}$	$\frac{m^+ - m_0}{m^+ - m_0 + \frac{1}{2}m^-}$	Dice
$S_{RT2}^{0,2}$	$\frac{m^+ - m_0}{m^+ - m_0 + 2 \cdot m^-}$	Rogers & Tanimoto II
S_{K2}	$\frac{m^+ - m_0}{m^-}$	Kulczynski II
S_{H2}	$\frac{m^+ - m_0 - m^-}{m^+ - m_0 + m^-}$	Hamman II

As in the binary case measures that can be derived from Eq. (4) are restricted to [0,1]. The measures of Hamman may result in negative similarities but do not exceed 1 whereas the coefficients of Kulczynski are nonnegative but may have values > 1.

Flexible Matching Coefficients

A first step to generalise the usual measures of agreement as given by Eq. (4) is to allow for $\lambda \geq 0$.

For instance, λ = ¾ yields a similarity measure called Quarterprop also investigated by Müller (2004). Thus, it is possible to include the 0-0-matches in the assessment of similarity but assigning them lower importance as the remaining matches.

In the particular situation of SNP data with 3 categories this leads to

$$S^{\lambda,\delta} = \frac{m_{22} + m_{11} + \lambda m_{00}}{m_{22} + m_{11} + \lambda m_{00} + \delta\left(m_{21} + m_{12} + m_{20} + m_{02} + m_{10} + m_{01}\right)},$$

with $\lambda \geq 0$ and $\delta > 0$.

The next step towards a generalisation is to permit different weights for different groups of matches and mismatches

$$S^{flex-ii,\lambda,\delta} = \frac{\lambda_2 m_{22} + \lambda_1 m_{11} + \lambda_0 m_{00}}{\lambda_2 m_{22} + \lambda_1 m_{11} + \lambda_0 m_{00} + \delta_{02}\left(m_{02} + m_{20}\right) + \delta_{01}\left(m_{01} + m_{10}\right) + \delta_{12}\left(m_{12} + m_{21}\right)} \quad (6)$$

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

Thus, it is possible to stress the importance of the least frequent 2-2-matches, i.e. common occurrence of a homozygous SNP, and to consider, for instance, homozygous references and heterozygous types as less different as homozygous references and variants. So, it is reasonable to assume that $\lambda_2 \geq \lambda_1 \geq \lambda_0 \geq 0$ stressing the importance of the common occurrence of homozygous variants and $\delta_{02} \geq \delta_{01} > 0$ and $\delta_{02} \geq \delta_{12} > 0$ so that homozygous variants and references are set to be most dissimilar.

A further extension consists in an extended definition of agreement. Assume that a common occurrence of at least one SNP is rather a similar genotype

combination than a dissimilar one. Thus, the respective numbers of combinations may be treated as matches but perhaps with a lower weight:

$$S^{flex-12,\lambda,\delta} = \frac{\lambda_2 m_{22} + \lambda_1 m_{11} + \lambda_0 m_{00} + \lambda_{12}(m_{12} + m_{21})}{\lambda_2 m_{22} + \lambda_1 m_{11} + \lambda_0 m_{00} + \lambda_{12}(m_{12} + m_{21}) + \delta_{02}(m_{02} + m_{20}) + \delta_{01}(m_{01} + m_{10})} \quad (7)$$

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

Similar to (7) the presence of at least one reference copy might be regarded as contributing to the similarity of two variables or subjects.

$$S^{flex-01,\lambda,\delta} = \frac{\lambda_2 m_{22} + \lambda_1 m_{11} + \lambda_0 m_{00} + \lambda_{01}(m_{01} + m_{10})}{\lambda_2 m_{22} + \lambda_1 m_{11} + \lambda_0 m_{00} + \lambda_{01}(m_{01} + m_{10}) + \delta_{02}(m_{02} + m_{20}) + \delta_{12}(m_{12} + m_{21})} \quad (8)$$

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

Eq. (6)-(8) can easily be generalised to the *p* categorical case and a further extension of the definition of matches and mismatches:

$$S^{flex-IJ,\lambda,\delta} = \frac{\sum_{i \in I} \lambda_i m_i}{\sum_{i \in I} \lambda_i m_i + \sum_{j \in J} \delta_j m_j},$$

where $\sum_{i \in I} \lambda_i m_i$ is the weighted sum of matches and $\sum_{j \in J} \delta_j m_j$ is the weighted sum of mismatches, *I* is the index set for similar categories and *J* is the index set for dissimilar categories. For convenience and to assure the symmetry of the corresponding similarity matrix for all variables or objects the indices *kl* and *lk* are pooled together to one index *kl*, $k \leq l$. Note that $m_{kl} = m'_{kl} + m'_{lk}$, $\forall k, l = 1, \ldots, p$, $k \leq l$, 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 $S^{flex\text{-}IJ,\lambda,\delta}$: $V \times V \rightarrow IR$, and $S^{flex\text{-}IJ,\lambda,\delta}$: $O \times O \rightarrow IR$, respectively, is given by

$$S^{flex-IJ,\lambda,\delta} := \frac{\Lambda}{\Lambda + \Delta}, \tag{9}$$

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

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

$J = \{j=kl, k \leq l, k, l = 0, 1, \ldots, p|$all combinations of category k and l 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, $\lambda_i \geq 0, \forall i \in I$, $\sum_{i \in I} \lambda_i > 0$, $\delta_j \geq 0, \forall j \in J$, $\sum_{j \in J} \delta_j > 0$, and $m_i \geq 0, \forall i \in I$, $m_j \geq 0, \forall j \in J$, $\sum_{i \in I} m_i + \sum_{j \in J} m_j > 0$ 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 Appendix.

REMARK 6. Special cases

Equations (6) – (8) are special cases of (9) with I = {0, 1, 2} and J = {02, 01, 12} for Eq. (6), I = {0, 1, 2, 12} and J = {02, 01} for Eq. (7) and I = {0, 1, 2, 01} and J = {02, 12} for Eq. (8).

In particular, these similarity measures satisfy (A1) – (A5).

Considering the function $S^{flex-IJ,\lambda,\delta}$ with respect to its dependence on the parameters λ_i and δ_j we use the following abbreviations:

$$\Lambda^{-i} := \Lambda - \lambda_i m_i = \sum_{i' \in I \setminus \{i\}} \lambda_{i'} m_{i'}, \quad \Lambda^{-i,i'} := \Lambda - \lambda_i m_i - \lambda_{i'} m_{i'} \tag{10}$$

$$\Delta^{-j} := \Delta - \delta_j m_j = \sum_{j' \in J \setminus \{j\}} \delta_{j'} m_{j'}, \; \Delta^{-j,j'} := \Delta - \delta_j m_j - \delta_{j'} m_{j'} \tag{11}$$

So, $S^{flex-IJ,\lambda,\delta}$ has the following properties.

THEOREM 1. Properties of $S^{flex-IJ,\lambda,\delta}(\lambda_i \mid V_k, V_l)$

Let $S^{flex-IJ,\lambda,\delta}$ be a measure of similarity as given by Definition 3 and let $\Lambda^{-i} + \Delta > 0$ and $m_i > 0$. Then $S^{flex-IJ,\lambda,\delta}(\lambda_i \mid V_k, V_l)$ has the following properties for all λ_i, $i \in I$.

i. $$S^{flex-IJ,\lambda,\delta}(\lambda_i = 0 \mid V_k, V_l) = \frac{\Lambda^{-i}}{\Lambda^{-i} + \Delta}$$

ii. $$\lim_{\lambda_i \to 0^+} S^{flex-IJ,\lambda,\delta}(\lambda_i \mid V_k, V_l) = \frac{\Lambda^{-i}}{\Lambda^{-i} + \Delta}$$

iii. $$\lim_{\lambda_i \to \infty} S^{flex-IJ,\lambda,\delta}(\lambda_i \mid V_k, V_l) = 1$$

iv. $$\frac{\partial}{\partial \lambda_i} S^{flex-IJ,\lambda,\delta}(\lambda_i \mid V_k, V_l) = \frac{m_i \Delta}{\left(\lambda_i m_i + \Lambda^{-i} + \Delta\right)^2} \geq 0$$

v. $$\frac{\partial}{\partial \lambda_i} S^{flex-IJ,\lambda,\delta}(\lambda_i = 0 \mid V_k, V_l) = \frac{m_i \Delta}{\left(\Lambda^{-i} + \Delta\right)^2} > 0 \text{, } \forall \; m_i \text{ and } \Delta > 0$$

vi. $$\frac{\partial}{\partial \lambda_i \partial \lambda_i} S^{flex-IJ,\lambda,\delta}(\lambda_i \mid V_k, V_l) = \frac{-2{m_i}^2 \Delta}{\left(\lambda_i m_i + \Lambda^{-i} + \Delta\right)^3} \leq 0$$

vii. $$\frac{\partial}{\partial \lambda_i \partial \lambda_i} S^{flex-IJ,\lambda,\delta}(\lambda_i = 0 \mid V_k, V_l) = \frac{-2{m_i}^2 \Delta}{\left(\Lambda^{-i} + \Delta\right)^3} < 0 \text{, } \forall \; m_i \text{ and } \Delta > 0.$$

PROOF: see Appendix.

This means that $S^{flex-IJ,\lambda,\delta}(\lambda_i \mid V_k, V_l)$ is a continuous monotonically increasing function approximating 1 for $\lambda_i \to \infty$ with minimum $S^{flex-IJ,\lambda,\delta}(0 \mid V_k, V_l) = \frac{\Lambda^{-i}}{\Lambda^{-i} + \Delta}$ but no inflexion point (see Figures 6 and 7 for illustration).

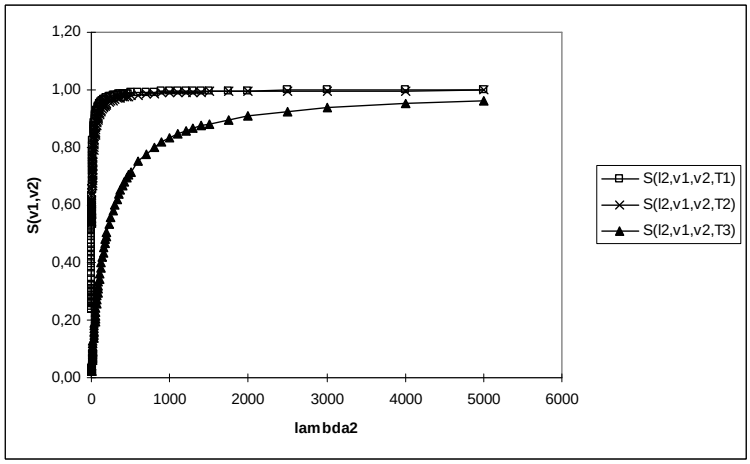

Figure 6. *$S^{flex-IJ,\lambda,\delta}(\lambda_2 \mid v_1, v_2)$ for different contingency tables T1, T2, and T3.*

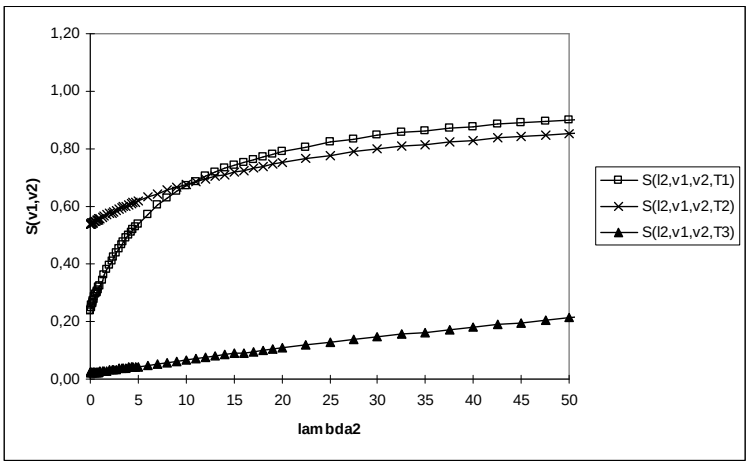

Figure 7. *$S^{flex-IJ,\lambda,\delta}(\lambda_2 \mid v_1, v_2)$ for different contingency tables T1, T2, and T3.*

The values of $S^{flex-IJ,\lambda,\delta}(\lambda_2 \mid v_1, v_2)$ are calculated using λ_1 = 1, λ_0 = 0.5, δ_{02} = 2, δ_{01} = δ_{12} = 1 and three different contingency tables T1, T2 and T3, where T1 represents a rather balanced contingency table with $m^+ = m_0+m_1+m_2 = 90+100+80 = 270$ and m^- $= m_{01}+m_{10}+m_{02}+m_{20}+m_{12}+m_{21} = 70+65+35+85+50+40 = 345$. Table T2 represents the current situation with the SNP data with $m^+ = 300+100+20 = 420$ and $m^-= 20+30+35+15+40+25 =165$ and table T3 is a rather balanced table with few matches $m^+ = 25+10+5 = 40$ and m^- $= 90+160+150+130+100+95= 725$ mismatches.

THEOREM 2. Properties of $S^{flex-IJ,\lambda,\delta}(\lambda_i, \lambda_{i'} \mid V_k, V_l)$

Let $S^{flex-IJ,\lambda,\delta}$ be a measure of similarity as given by Definition 3 and let $\Lambda^{-i,i'} + \Delta > 0$ and $m_i + m_{i'} > 0$. Then $S^{flex-IJ,\lambda,\delta}(\lambda_i, \lambda_{i'} \mid V_k, V_l)$ has the following properties for all λ_i, $\lambda_{i'}$, i, $i' \in I$, $i \neq i'$:

i. $S^{flex-IJ,\lambda,\delta}(0,0 \mid V_k, V_l) = \dfrac{\Lambda^{-i,i'}}{\Lambda^{-i,i'} + \Delta}$

ii. $\lim\limits_{\substack{\lambda_i \to 0^+ \\ \lambda_{i'} \to 0^+}} S^{flex-IJ,\lambda,\delta}(\lambda_i, \lambda_{i'} \mid V_k, V_l) = \dfrac{\Lambda^{-i,i'}}{\Lambda^{-i,i'} + \Delta}$

iii. $\lim\limits_{\substack{\lambda_i \to \infty \\ \lambda_{i'} \to \infty}} S^{flex-IJ,\lambda,\delta}(\lambda_i, \lambda_{i'} \mid V_k, V_l) = 1.$

PROOF: see Appendix.

So $S^{flex-IJ,\lambda,\delta}(\lambda_i, \lambda_{i'} \mid V_k, V_l)$ is continuous in $(\lambda_i,\lambda_{i'}) = (0,0)$ with minimum $S^{flex-IJ,\lambda,\delta}(0,0 \mid V_k, V_l) = \dfrac{\Lambda^{-i,i'}}{\Lambda^{-i,i'} + \Delta}$ and approximates 1 for $\lambda_i \to \infty$, $\lambda_{i'} \to \infty$.

Now, we consider the dependence of $S^{flex-IJ,\lambda,\delta}$ on the weights of the mismatches δ_j.

THEOREM 3. Properties of $S^{flex-IJ,\lambda,\delta}(\delta_j \mid V_k, V_l)$

Let $S^{flex-IJ,\lambda,\delta} : IR_0^+ \to [0,1]$ be a measure of similarity as given by Definition 3 and let $\Lambda + \Delta^{-j} > 0$ and $m_j > 0$. Then $S^{flex-IJ,\lambda,\delta}(\delta_j \mid V_k, V_l)$ has the following properties for all δ_j, $j \in J$.

i. $S^{flex-IJ,\lambda,\delta}(0 \mid V_k, V_l) = \dfrac{\Lambda}{\Lambda + \Delta^{-j}}$

ii. $\lim\limits_{\delta_j \to 0^+} S^{flex-IJ,\lambda,\delta}(\delta_j \mid V_k, V_l) = \dfrac{\Lambda}{\Lambda + \Delta^{-j}}$

iii. $\lim\limits_{\delta_j \to \infty} S^{flex-IJ,\lambda,\delta}(\delta_j \mid V_k, V_l) = 0$

iv. $$\frac{\partial}{\partial \delta_j} S^{flex-IJ,\lambda,\delta}(\delta_j \mid V_k, V_l) = \frac{-\Lambda m_j}{\left(\Lambda + \Delta^{-j} + \delta_j m_j\right)^2} \leq 0$$

v. $$\frac{\partial}{\partial \delta_j} S^{flex-IJ,\lambda,\delta}(\delta_j = 0 \mid V_k, V_l) = \frac{-\Lambda m_j}{\left(\Lambda + \Delta^{-j}\right)^2} < 0 \text{, } \forall\ m_j \text{ and } \Lambda > 0$$

vi. $$\frac{\partial}{\partial \delta_j \partial \delta_j} S^{flex-IJ,\lambda,\delta}(\delta_j \mid V_k, V_l) = \frac{2\Lambda m_j^{\,2}}{\left(\Lambda + \Delta^{-j} + \delta_j m_j\right)^3} \geq 0$$

vii. $$\frac{\partial}{\partial \delta_j \partial \delta_j} S^{flex-IJ,\lambda,\delta}(\delta_j = 0 \mid V_k, V_l) = \frac{2\Lambda m_j^{\,2}}{\left(\Lambda + \Delta^{-j}\right)^3} > 0 \text{, } \forall\ m_j \text{ and } \Lambda > 0.$$

PROOF: see Appendix.

This means that $S^{flex-IJ,\lambda,\delta}(\delta_j \mid V_k, V_l)$ is a continuous monotonically decreasing function approximating 0 for $\delta_j \to \infty$ with maximum $S^{flex-IJ,\lambda,\delta}(0 \mid V_k, V_l) = \frac{\Lambda}{\Lambda + \Delta^{-j}}$ but no inflexion point (see Figure 8 for illustration).

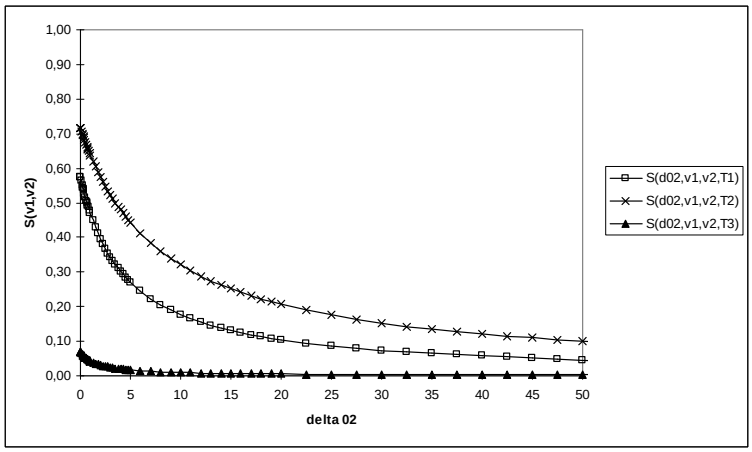

Figure 8. *$S^{flex-IJ,\lambda,\delta}(\delta_{02} \mid v_1, v_2)$ for different contingency tables T1, T2, and T3.*

The values of $S^{flex-IJ,\lambda,\delta}(\delta_{02} \mid v_1, v_2)$ are calculated using $\lambda_2 = 2$, $\lambda_1 = 1$, $\lambda_0 = 0.5$, $\delta_{01} = \delta_{12} = 1$ and three different contingency tables T1, T2 and T3 as described above.

THEOREM 4. **Properties of $S^{flex-IJ,\lambda,\delta}(\delta_j, \delta_{j'} \mid V_k, V_l)$**

Let $S^{flex-IJ,\lambda,\delta} : IR_0^+ \times IR_0^+ \to [0,1]$ be a measure of similarity as given by Definition 3 and let $\Lambda + \Delta^{-j,j'} > 0$ and $m_j + m_{j'} > 0$. Then $S^{flex-IJ,\lambda,\delta}(\delta_j, \delta_{j'} \mid V_k, V_l)$ has the following properties for all δ_j, $\delta_{j'}$, $j, j' \in J$, $j \neq j'$.

i. $$S^{flex-IJ,\lambda,\delta}(0, 0 \mid V_k, V_l) = \frac{\Lambda}{\Lambda + \Delta^{-j,j'}}$$

ii. $$\lim_{\substack{\delta_j \to 0^+ \\ \delta_j \to 0^+}} S^{flex-IJ,\lambda,\delta}(\delta_j, \delta_{j'} \mid V_k, V_l) = \frac{\Lambda}{\Lambda + \Delta^{-j,j'}}$$

iii. $$\lim_{\substack{\delta_j \to \infty \\ \delta_j \to \infty}} S^{flex-IJ,\lambda,\delta}(\delta_j, \delta_{j'} \mid V_k, V_l) = 0 .$$

PROOF: see Appendix.

So $S^{flex-IJ,\lambda,\delta}(\delta_j, \delta_{j'} \mid V_k, V_l)$ approximates $\frac{\Lambda}{\Lambda + \Delta^{-j,j'}}$ in $(\delta_j, \delta_{j'}) = (0,0)$ and approximates 0 for $\delta_j \to \infty$, $\delta_{j'} \to \infty$.

Now, we consider the joint dependence of $S^{flex\ -IJ,\lambda,\delta}$ on the weights of the matches λ_i and on the weights of the mismatches δ_j.

THEOREM 5. Properties of $S^{flex-IJ,\lambda,\delta}(\lambda_i,\delta_j \mid V_k,V_l)$

Let $S^{flex-IJ,\lambda,\delta}: IR_0^+ \times IR_0^+ \to [0,1]$ be a measure of similarity as given by Definition 3 and let $\Lambda^{-i} + \Delta^{-j} > 0$ and $m_i + m_j > 0$. Then $S^{flex-IJ,\lambda,\delta}(\lambda_i,\delta_j \mid V_k,V_l)$ has the following properties for all λ_i, $i \in I$, δ_j, $j \in J$.

i. $S^{flex-IJ,\lambda,\delta}(0,0 \mid V_k,V_l) = \dfrac{\Lambda^{-i}}{\Lambda^{-i} + \Delta^{-j}}$

ii. $\lim\limits_{\substack{\lambda_i \to 0^+ \\ \delta_j \to 0^+}} S^{flex-IJ,\lambda,\delta}(\lambda_i,\delta_j \mid V_k,V_l) = \dfrac{\Lambda^{-i}}{\Lambda^{-i} + \Delta^{-j}}$

iii. $0 = \lim\limits_{\delta_j \to \infty} S^{flex-IJ,\lambda,\delta}(\delta_j \mid V_k,V_l) \leq$

$$\lim_{\substack{\lambda_i \to \infty \\ \delta_j \to \infty}} S^{flex-IJ,\lambda,\delta}(\lambda_i,\delta_j \mid V_k,V_l) \leq \lim_{\lambda_i \to \infty} S^{flex-IJ,\lambda,\delta}(\lambda_i \mid V_k,V_l) = 1$$

PROOF: see Appendix.

So, $S^{flex-IJ,\lambda,\delta}(\lambda_i,\delta_j \mid V_k,V_l) \in [0, 1]$ approximates $\dfrac{\Lambda^{-i}}{\Lambda^{-i} + \Delta^{-j}}$. in $(\lambda_i,\delta_j) = (0,0)$. Figure 9 illustrates the behaviour of $S^{flex-IJ,\lambda,\delta}$ depending on λ_2 and δ_{02} for the contingency table T1.

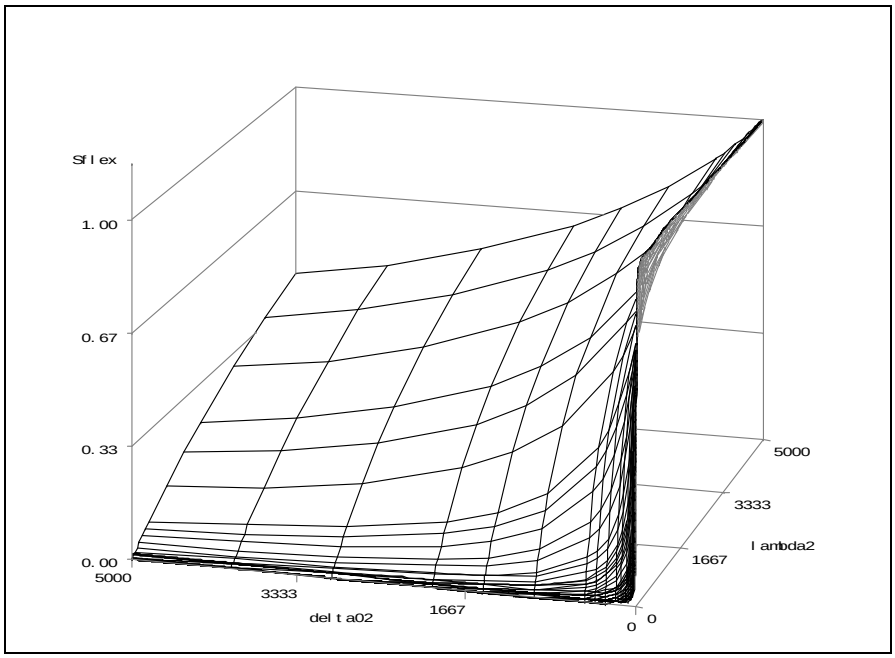

Figure 9. *$S^{flex-IJ,\lambda,\delta}(\lambda_2, \delta_{02} \mid v_1, v_2)$ for contingency table T1.*

The values of $S^{flex-IJ,\lambda,\delta}(\lambda_2, \delta_{02} \mid v_1, v_2)$ are calculated using $\lambda_2 = 2$, $\lambda_1 = 1$, $\lambda_0 = 0.5$, $\delta_{02} = 2$, $\delta_{01} = \delta_{12} = 1$ and contingency table T1 as described above.

Measures of dependence

In case of nominally 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 differently (Anderberg, 1973, Hartung, 1991).

We consider here Pearson's Corrected Coefficient of Contingency

$$S_{PC} = \sqrt{\frac{\min(p,q)}{\min(p,q)-1}} \cdot \sqrt{\frac{\chi^2}{\chi^2 + m}}, \qquad (12)$$

where p and q are the numbers of categories of the variables or objects, $m = \sum_{i=1}^{p} \sum_{j=1}^{p} m_{ij}$ is the total 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* on the table size.

Members of this class of measures include also Cramèr's C (see for example Müller *et al.*, 2005).

Pearson's Corrected Coefficient of Contingency is a useful tool to compare categorial variables. It allows for different numbers of categories and we are able to compare variables which are not similar be nature, 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.

4.3.2 Ordinal scale

In case of ordinal scaled data we can assess the proximity of two variables or objects 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 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.

Considering a 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 special 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 we have only three possible categories for 1200 observations comparing the variables or three possible observations 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 three categories that can be ordered and if the size of the tied groups is not too big.

5. 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).

We display here a selection of dendrograms to illustrate the effect of the choice of parameters and index sets of the Flexible Matching Coefficients as given by Definition 3. In particular we consider the special case of Equation 6 with $I = \{0, 1, 2\}$ and $J = \{02, 01, 12\}$, Equation 7 with $I = \{0, 1, 2, 12\}$ and $J = \{02, 01\}$ and Equation 8 with $I = \{0, 1, 2, 01\}$ and $J = \{02, 12\}$ for clustering variables.

Figure 10 to 22 result from different index sets and choices of parameters. Figures 13, 14, 18 and 19 show the results for cases and controls. See also Table 8 for an overview.

Table 8. *Case-control status and parameters of Fig. 12 - 24. Eq. denotes the respective equation.*

Figure	status	λ_2	λ_1	λ_0	λ_{12}	λ_{01}	δ_{12}	δ_{02}	δ_{01}	Eq.
10	control	1	1	0.5	-	-	0.5	1	1	(6)
11	control	2	1	0.5	-	-	0.5	1	0.5	(6)
12	control	2	1	0	-	-	0.5	2	1	(6)
13	case	2	1	0.66	-	-	0.33	1	0.33	(6)
14	control	2	1	0.66	-	-	0.33	1	0.33	(6)
15	control	1	1	0	1	-	-	1	1	(7)
16	control	1	1	1	1	-	-	1	1	(7)
17	control	2	1	0.5	0.5	-	-	2	1	(7)
18	case	2	1	0.66	0.33	-	-	2	1	(7)
19	control	2	1	0.66	0.33	-	-	2	1	(7)
20	control	1	1	0	-	1	1	1	-	(8)
21	control	1	1	1	-	1	1	1	-	(8)
22	control	4	1	0.5	-	0.5	1	4	-	(8)

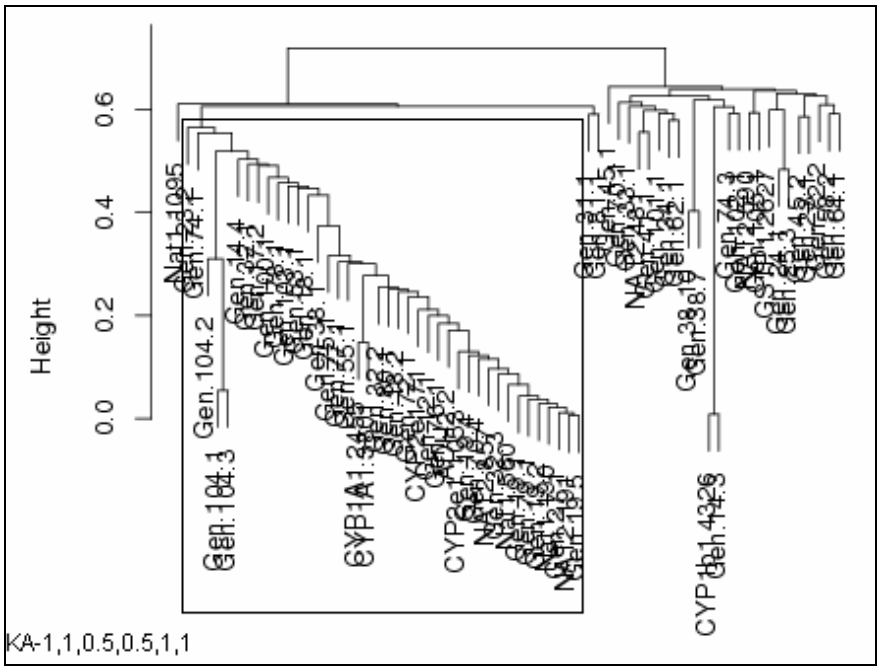

Figure 10. *Dendrogram of the flexible matching coefficients of the control group with I = {2, 1, 0}, J = {12, 02, 01}, λ = (1, 1, 0.5), δ = (0.5, 1, 1).*

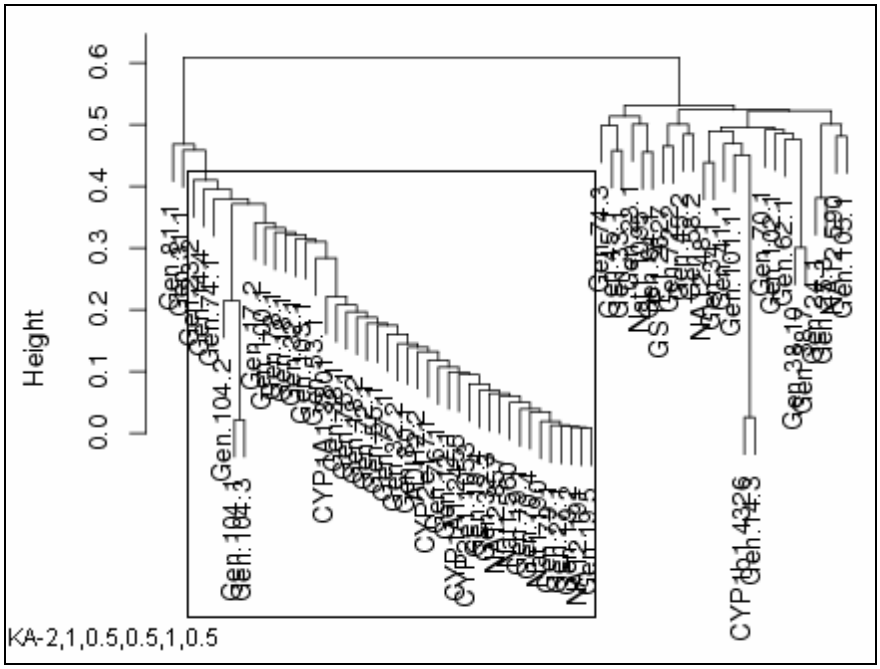

Figure 11. *Dendrogram of the flexible matching coefficients of the control group with I = {2, 1, 0}, J = {12, 02, 01}, λ = (2, 1, 0.5), δ = (0.5, 1, 0.5).*

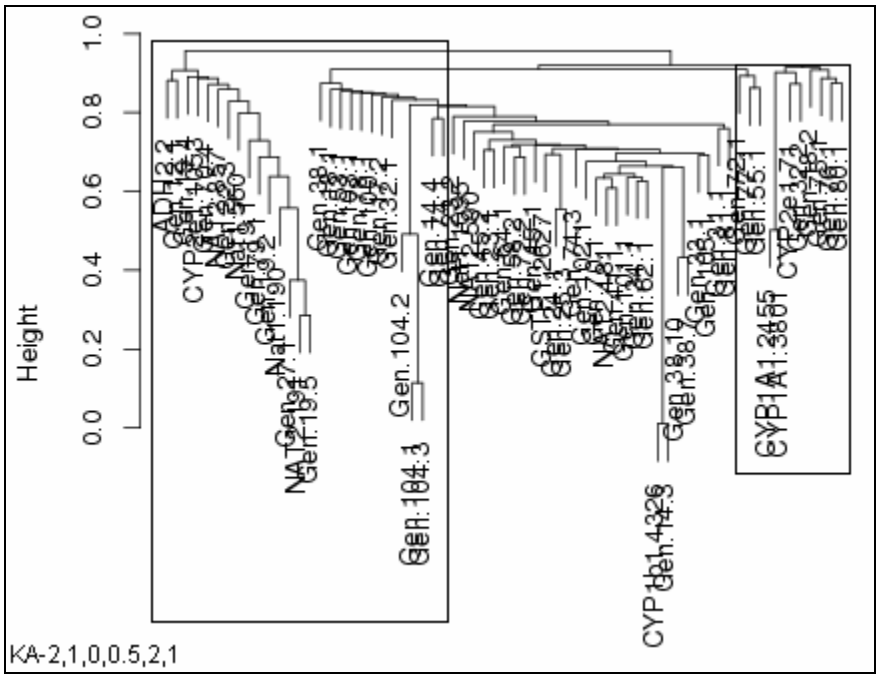

Figure 12. *Dendrogram of the flexible matching coefficients of the control group with I = {2, 1, 0), J = {12, 02, 01}, λ = (2, 1, 0), δ = (0.5, 2, 1).*

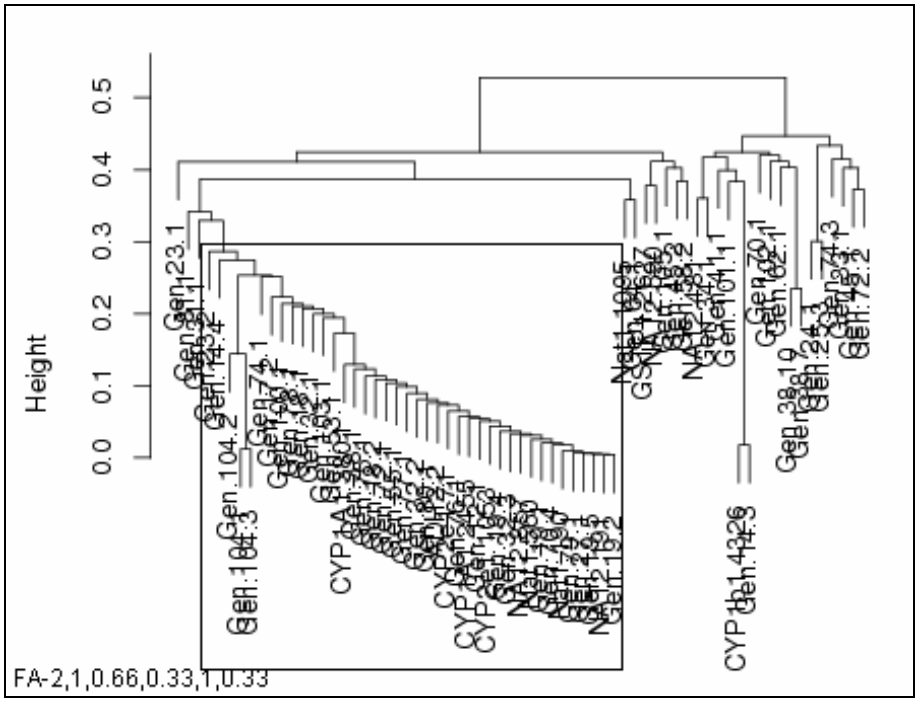

Figure 13. *Dendrogram of the flexible matching coefficients of the case group with I ={2, 1, 0}, J ={12, 02, 01}, λ =(2, 1, 0.66), δ =(0.33, 1, 0.33).*

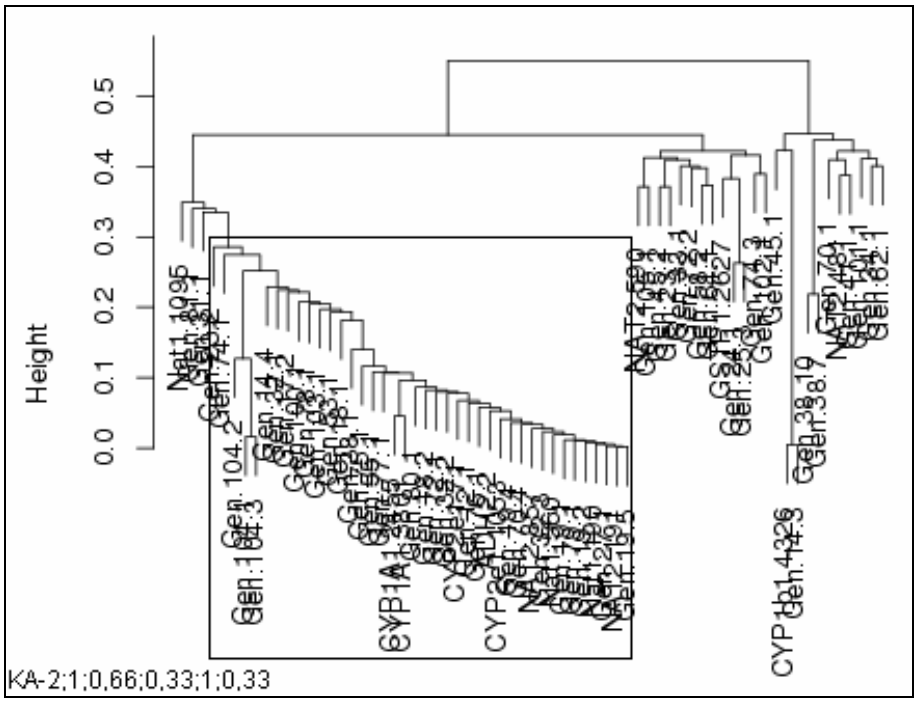

Figure 14. *Dendrogram of the flexible matching coefficients of the control group with I = {2, 1, 0), J = {12, 02, 01}, λ =(2, 1, 0.66), δ =(0.33, 1, 0.33).*

Considering the 1-2-combinations as matches and applying the coefficients of Jaccard, i.e. excluding the 0-0-matches, and Simple Matching , i.e. including the 0-0-matches, leads to Figures 15 and 16.

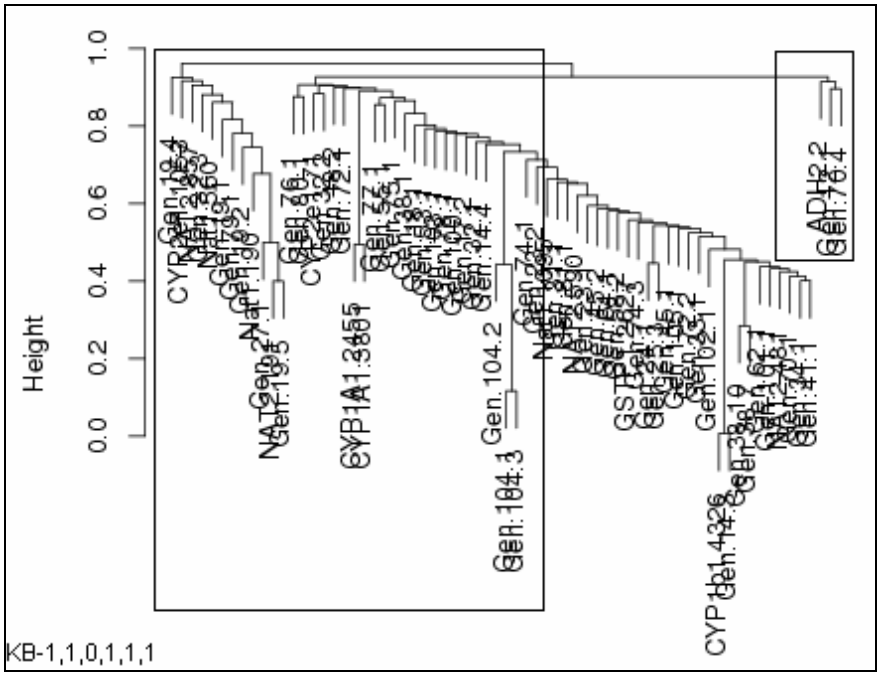

Figure 15. *Dendrogram of the flexible matching coefficients of the control group with I = {2, 1, 0, 12}, J = {02, 01}, λ =(1, 1, 0, 1), δ =(1, 1).*

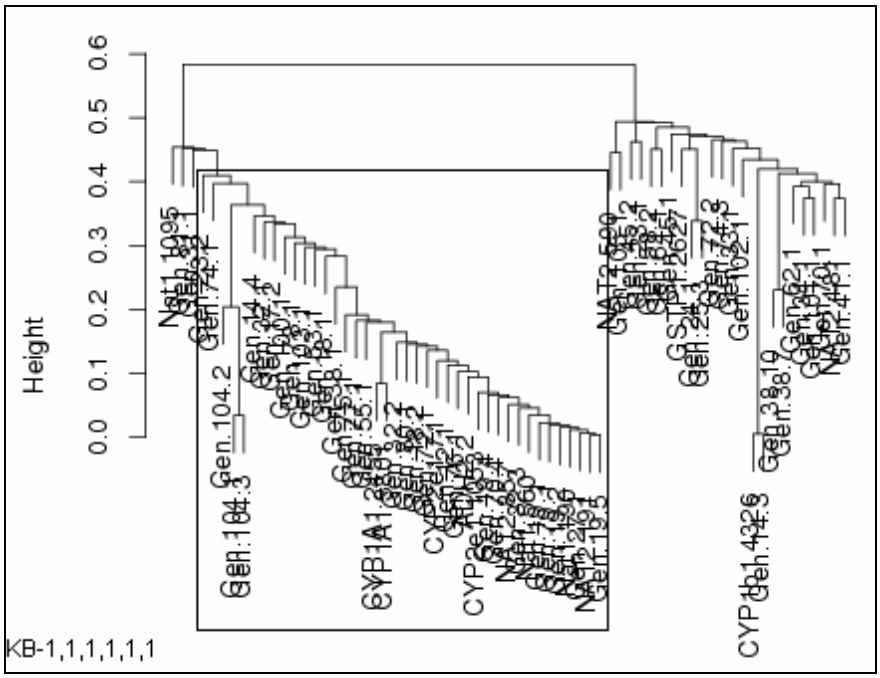

Figure 16. *Dendrogram of the flexible matching coefficients of the control group with I = {2, 1, 0, 12}, J = {02, 01}, λ =(1, 1, 1, 1), δ =(1, 1).*

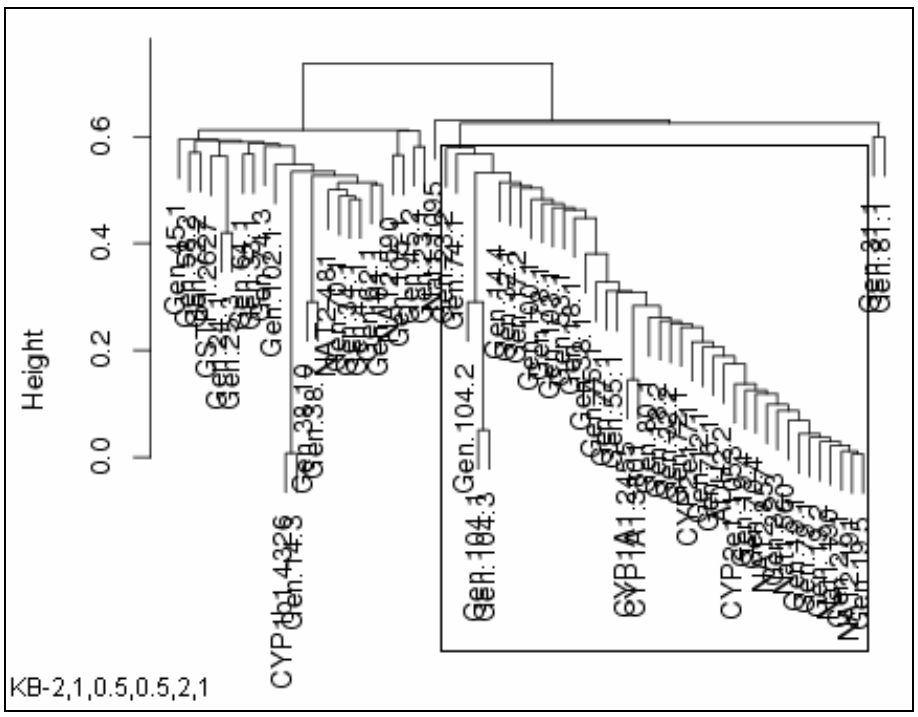

Figure 17. *Dendrogram of the flexible matching coefficients of the control group with I = {2, 1, 0, 12}, J = {02, 01}, λ =(2, 1, 0.5, 0.5), δ =(2, 1).*

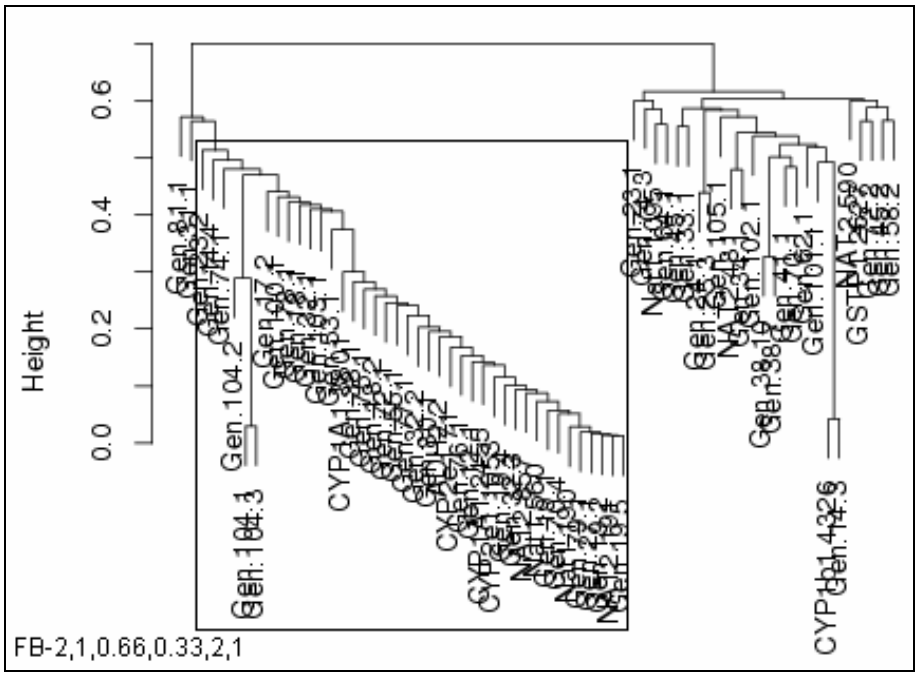

Figure 18. *Dendrogram of the flexible matching coefficients of the case group with I = {2, 1, 0, 12}, J = {02, 01}, λ =(2, 1, 0.66, 0.33), δ =(2, 1).*

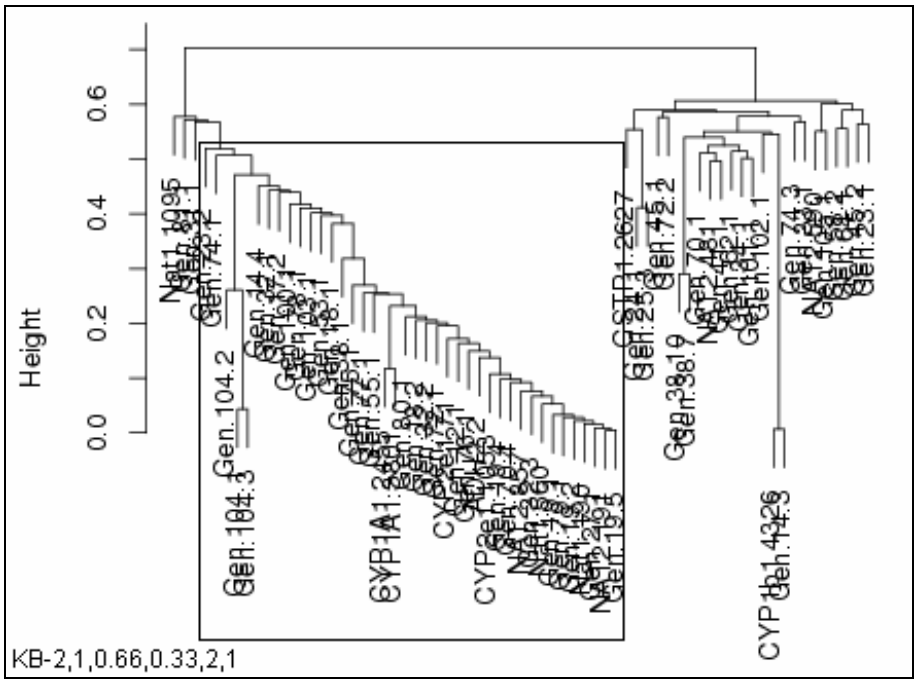

Figure 19. *Dendrogram of the flexible matching coefficients of the control group with I = {2, 1, 0, 12}, J = {02, 01}, λ =(2, 1, 0.66, 0.33), δ =(2, 1).*

Considering the 0-1-combinations as matches and applying the coefficients of Jaccard, i.e. excluding the 0-0-matches, and Simple Matching , i.e. including the 0-0-matches, leads to Figures 20 and 21.

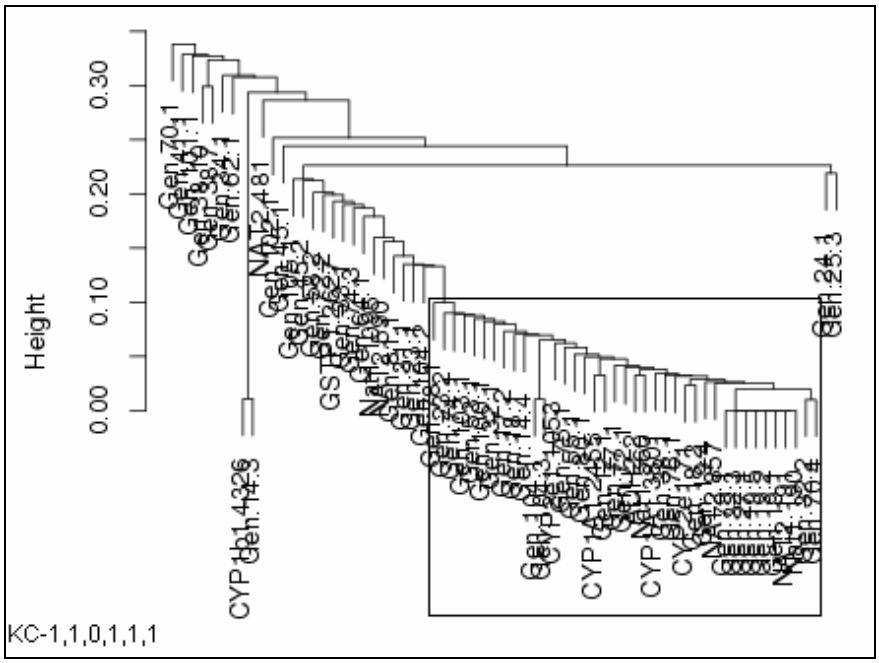

Figure 20. *Dendrogram of the flexible matching coefficients of the control group with I = {2, 1, 0, 01}, J = {01, 02}, λ =(1, 1, 0, 1), δ =(1, 1).*

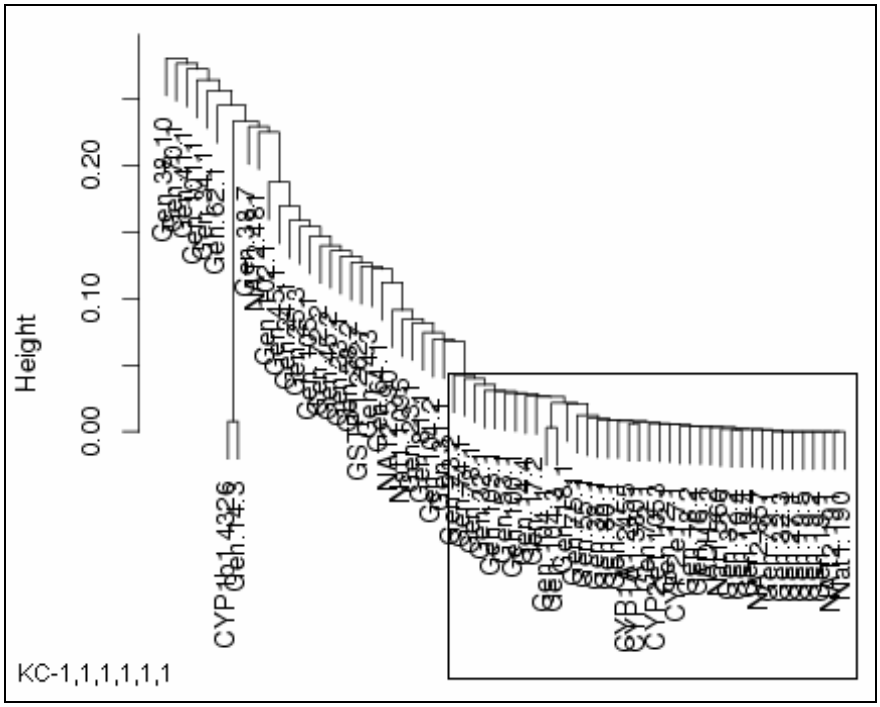

Figure 21. *Dendrogram of the flexible matching coefficients of the control group with I = {2, 1, 0, 01}, J = {01, 02}, λ =(1, 1, 1, 1), δ =(1, 1).*

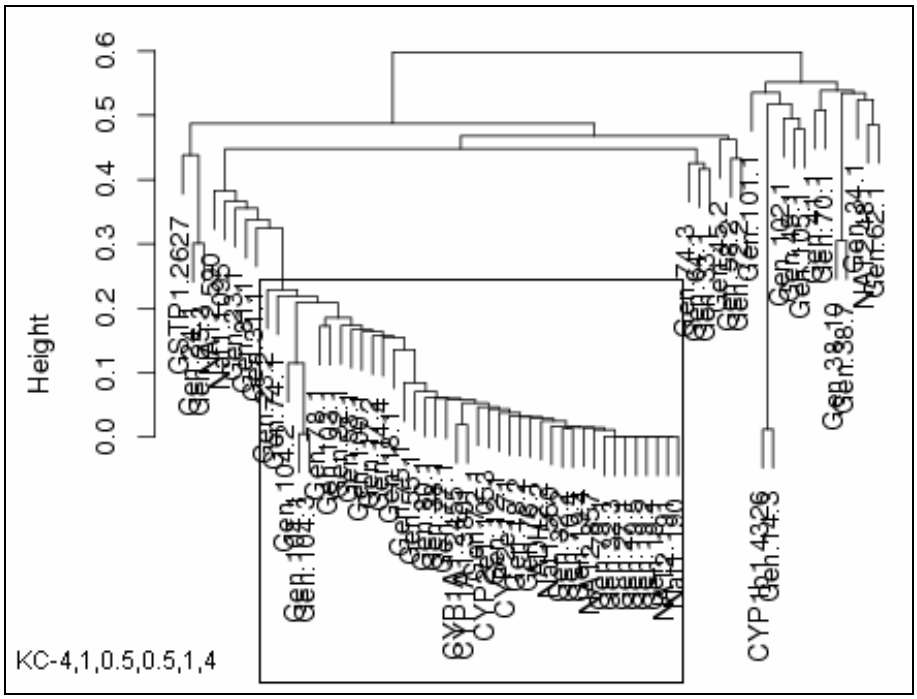

Figure 22. *Dendrogram of the flexible matching coefficients of the control group with I = {2, 1, 0, 01}, J = {01, 02}, λ =(4, 1, 0.5, 0.5), δ =(1, 4).*

5.1 Conclusions

Summarising the results for the conventional and the new matching coefficients as well as for measures based on the χ^2-statistic the usual matching coefficients form two groups depending on their consideration or ignorance of the 0-0-matches. Within each group the weight δ of the mismatches is of minor importance and has no impact on the structure of the dendrogram. For the present data set these measures yield poorly structured dendrograms similar to the results for $I = \{2, 1, 0, 01\}$ and $J = \{12, 02\}$ as shown in figures 20 and 21 where subgroups of variables cannot be detected and the dendrograms have the form of a stair resulting from the addition of one variable after the other to the sole big cluster. As shown in figures 10-22 Flexible Matching Coefficients yield more structured dendrograms as, for instance, Figures 13, 14 (cases, controls) and 18, 19 (cases, controls) with $I = \{2, 1, 0\}$, $J = \{12, 02, 01\}$, $\lambda = \{2, 1, 0.66\}$, $\delta = \{0.33, 1, 0.33\}$ and $I = \{2, 1, 0, 12\}$, $J = \{02, 01\}$, $\lambda = \{2, 1, 0.66, 0.33\}$, $\delta = \{2, 1\}$, respectively. The weights for the matches and mismatches have a small but clear impact on the clustering though the general structure remains unless the 0-0-matches are not excluded from the analysis (figures 12 and 15). Comparing the clustering from a number of variations of I, J, λ and δ a stable group of loci can be identified that shows minor variations between the different matching coefficients in cases as well as in controls. This group can also be found using the Corrected Contingency Coefficient of Pearson. Hence, this particular group of variables may be neglected for a further analysis. Applying a further cluster analysis to the remaining variables enables more insight into the differences between cases and controls.

Classification procedures may use representatives of the stable group instead of all of them reducing the amount of competing models.

Furthermore several small groups of two or three loci, some of the same, others of different genes, appear independently from the applied measure of similarity, for instant, the three investigated loci 1, 2 and 3 of gene 104, two of the three investigated loci 7 and 10 of gene 38 as well as the loci 24.1 and 25.3.

The general problem with all measures of similarity based on the χ^2-statistic occurs if the contingency table of two variables contains empty lines or columns so that one of the variables is treated as a constant. This may happen, for instance, if the data set contains monomorphic SNPs or if all variants of one variable are compared to the missing values of the other one.

6. Discussion

The present approach is a promising tool to detect a general structure in SNP data as well as to find potential differences between cases and controls, i.e. variables and especially groups of variables that might be relevant for the assumed differences between cases and controls. A more detailed comparison of the conventional matching coefficients, further similarity coefficients and specific Flexible Matching Coefficients is presented in Müller *et al.* (2005) and Müller (2004). The addition of new variables seems to have minor impact on the general structure of the dendrogram so that the applied measures seem to result in a conserved structure. The latter can also

be observed considering the clusterings resulting from Pearson's Corrected Coefficient of Contingency.

In section 4.1 we discuss the development of measures which enable a comparison of the genotypes at the investigated loci with respect to their impact on the metabolism. As the effect or – more likely – multiple effects of each SNP plus synergistic effects of several SNPs, of the same gene, for instance, remain elusive for most of the considered loci, the development of such comparable measures of effects of point mutations remains a matter of future research.

The cluster analysis presented in Chapter 5 concentrates on the comparison of variables. The comparison of persons is omitted here as we focus on the performance of the similarity measures. Such an attempt that is rather difficult considering over 1200 subjects. There it is difficult to detect subgroups and structures in the resulting dendrograms using only the genetic variables.

Due to the nature of the problem we cannot restrict the analysis solely to the SNP data but have to account for further, exogenous factors. So the next step is a joint analysis of SNP data and exogenous risk or beneficial factors. This raises the problem of appropriate measures for different types of data, especially of differently scaled variables. This aspect is considered in detail by Selinski (2005) where mixed measures for clustering subjects as well as mixed measures and strategies for clustering variables of different scale and interpretation are presented.

In general, cluster analysis can help to gain insight into the data but especially in complex data sets it is reasonably combined with further approaches. For the detection of interactions between gene loci and 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 for SNP data is usually 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 –help to gain more insight and to develop biological hypotheses.

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.

Brazma A, Vilo J (2000). Gene expression data analysis. *FEBS Letters* **480**, pp. 17-24.

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.

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.

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.

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

Selinski S (2005). Similarity measures for clustering SNP and epidemiological data. *Technical Report*, University of Dortmund (*in prep.*).

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

Snustad DP and Simmons MJ (1999). *Principles of genetics.* 2nd ed., Wiley, New York.

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

Thier R, Brüning T, Roos PH, Rihs HP, Golka K, Ko Y and Bolt HM (2003). Markers of genetic susceptibility in human environmental hygiene and toxicology: the role of selected CYP, NAT and GST genes. *Int. J. Hyg. Environ. Health* **206**, pp. 149-71.

Tibshirani R, Walther G, Hastie T (2001). Estimating the number of clusters in a dataset via the gap statistic. *J. Royal Stat. Soc. B* **63**, pp. 411-423.

Zhang B and 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).*

ZTG Zentrum für Telematik im Gesundheitswesen GmbH (2004). Landesgesundheitsportal NRW – Brustkrebs. www. gesundheit.nrw.de.

Appendix

PROOF of REMARK 5:

Let m_{ikl}, m_{ikk} and m_{ikm} be the values of the matching category $i \in I$ of the variables V_k, V_l , V_k, V_k and V_k, V_m respectively. Let m_{jkl}, m_{jkk} and m_{jkm} be the values of mismatching categories $j \in J$, of the variables V_k, V_l, V_k, V_k and V_k, V_m respectively. Furthermore, let $\Lambda_{kl} := \sum_{i \in I} \lambda_i m_{ikl}$ and $\Delta_{kl} := \sum_{j \in J} \delta_j m_{jkl}$. The terms Λ_{kk}, Λ_{km}, Δ_{kk} and Δ_{km} are defined analogous.

To proof (A1) assume that V_k is more similar to V_l than to V_m, $V_l \neq V_m$. Hence, $\sum_{i \in I} \lambda_i m_{ikl} \geq \sum_{i \in I} \lambda_i m_{ikm}$ and $\sum_{j \in J} \delta_j m_{jkl} \leq \sum_{j \in J} \delta_j m_{jkm}$ with at least one inequality. Therefore

$$\Rightarrow S(V_k, V_l) = \frac{\Lambda_{kl}}{\Lambda_{kl} + \Delta_{kl}} = \frac{1}{1 + \Delta_{kl}/\Lambda_{kl}} > \frac{1}{1 + \Delta_{kl}/\Lambda_{km}} > \frac{1}{1 + \Delta_{km}/\Lambda_{km}}$$
$$= \frac{\Lambda_{km}}{\Lambda_{km} + \Delta_{km}} = S(V_k, V_m).$$

To proof (A2) is true as consider $S(V_k, V_l) = S(V_l, V_k)$

$$\Leftrightarrow \frac{\sum_{i \in I} \lambda_i m_{ikl}}{\sum_{i \in I} \lambda_i m_{ikl} + \sum_{j \in J} \delta_j m_{jkl}} = \frac{\sum_{i \in I} \lambda_i m_{ilk}}{\sum_{i \in I} \lambda_i m_{ilk} + \sum_{j \in J} \delta_j m_{jlk}}.$$

as $m_{ikl}=m_{ilk}$, $m_{jkl}=m_{jlk}$, $\forall\ i \in I, j \in J$.

To proof (A3) consider V_k, $V_l \in V$. Then

$$S(V_k, V_k) = \frac{\sum_{i \in I} \lambda_i m_{ikk}}{\sum_{i \in I} \lambda_i m_{ikk} + \sum_{j \in J} \delta_j m_{jkk}} = \frac{\Lambda_{kk}}{\Lambda_{kk}} = 1$$
$$\geq \frac{\sum_{i \in I} \lambda_i m_{ikl}}{\sum_{i \in I} \lambda_i m_{ikl} + \sum_{j \in J} \delta_j m_{jkl}} = \frac{\Lambda_{kl}}{\Lambda_{kl} + \Delta_{kl}} = S(V_k, V_l)$$

as $\Delta_{kl} \geq 0 \; \forall \; V_k, V_l \in V$.

(A4) $S(V_k, V_l) \geq 0$ is true as $\lambda_i \geq 0$, $\delta_j \geq 0$, $m_{ikl} \geq 0$ and $m_{jkl} \geq 0$ by definition and so $\Lambda_i \geq 0$, $\forall \; i \in I$, and $\Delta_j \geq 0$, $\forall \; j \in J$. Hence,

$$S(V_k, V_l) = \frac{\sum_{i \in I} \lambda_i m_{ikl}}{\sum_{i \in I} \lambda_i m_{ikl} + \sum_{j \in J} \delta_j m_{jkl}} = \frac{\Lambda_{kl}}{\Lambda_{kl} + \Delta_{kl}} \geq 0 \,, \forall \; V_k, V_l \in V.$$

To proof (A5) recall that $\Delta_{kk} = 0 \; \forall \; k$. Hence,

$$S(V_k, V_k) = \frac{\sum_{i \in I} \lambda_i m_{ikk}}{\sum_{i \in I} \lambda_i m_{ikk} + \sum_{j \in J} \delta_j m_{jkk}} = \frac{\Lambda_{kk}}{\Lambda_{kk}} = 1, \; \forall \; V_k, \in V.$$

❐

PROOF of THEOREM 1:

i., iv. – vii. trivial

ii. $$\lim_{\lambda_i \to 0^+} S^{flex-IJ,\lambda,\delta}(\lambda_i \,|\, V_k, V_l) = \lim_{\lambda_i \to 0^+} \frac{\Lambda^{-i} + \lambda_i m_i}{\Lambda^{-i} + \Delta + \lambda_i m_i} = \frac{\Lambda^{-i}}{\Lambda^{-i} + \Delta}$$

iii. $$\lim_{\lambda_i \to \infty} S^{flex-IJ,\lambda,\delta}(\lambda_i \,|\, V_k, V_l) = \lim_{\lambda_i \to \infty} \frac{\Lambda^{-i} + \lambda_i m_i}{\Lambda^{-i} + \Delta + \lambda_i m_i} \underset{\text{l'Hospital}}{=} \frac{\lim_{\lambda_i \to \infty} m_i}{\lim_{\lambda_i \to \infty} m_i} = \frac{m_i}{m_i} = 1$$

❐

PROOF of THEOREM 2.

i. trivial

ii. $$\lim_{\substack{\lambda_i \to 0^+ \\ \lambda_{i'} \to 0^+}} S^{flex-IJ,\lambda,\delta}(\lambda_i, \lambda_{i'} \,|\, V_k, V_l) = \lim_{\substack{\lambda_i \to 0^+ \\ \lambda_{i'} \to 0^+}} \frac{\Lambda^{-i,i'} + \lambda_i m_i + \lambda_{i'} m_{i'}}{\Lambda^{-i,i'} + \Delta + \lambda_i m_i + \lambda_{i'} m_{i'}} = \frac{\Lambda^{-i,i'}}{\Lambda^{-i,i'} + \Delta}$$

iii. Let α, β and $n > 0$, a and $b \geq 0$. Then,

$$\lim_{\substack{\lambda_i \to \infty \\ \lambda_{i'} \to \infty}} S^{flex-IJ,\lambda,\delta}(\lambda_i, \lambda_{i'} \mid V_k, V_l) = \lim_{\substack{\lambda_i \to \infty \\ \lambda_{i'} \to \infty}} \frac{\Lambda^{-i,i'} + \lambda_i m_i + \lambda_{i'} m_{i'}}{\Lambda^{-i,i'} + \Delta + \lambda_i m_i + \lambda_{i'} m_{i'}}$$

$$= \lim_{n \to \infty} \frac{\Lambda^{-i,i'} + (\alpha n + a) m_i + (\beta n + b) m_{i'}}{\Lambda^{-i,i'} + \Delta + (\alpha n + a) m_i + (\beta n + b) m_{i'}} \overset{\text{l'Hospital}}{=} \frac{\lim_{n \to \infty} \alpha m_i + \beta m_{i'}}{\lim_{n \to \infty} \alpha m_i + \beta m_{i'}} = 1$$

❐

PROOF of THEOREM 3

i., iv. – vii. trivial

ii. $$\lim_{\delta_j \to 0^+} S^{flex-IJ,\lambda,\delta}(\delta_j \mid V_k, V_l) = \lim_{\delta_j \to 0^+} \frac{\Lambda}{\Lambda + \Delta^{-j} + \delta_j m_j} = \frac{\Lambda}{\Lambda + \Delta^{-j}}$$

iii. $$\lim_{\delta_j \to \infty} S^{flex-IJ,\lambda,\delta}(\delta_j \mid V_k, V_l) = \lim_{\delta_j \to \infty} \frac{\Lambda}{\Lambda + \Delta^{-j} + \delta_j m_j} = 0 .$$

❐

PROOF of THEOREM 4

i. trivial

ii. $$\lim_{\substack{\delta_j \to 0^+ \\ \delta_j \to 0^+}} S^{flex-IJ,\lambda,\delta}(\delta_j, \delta_{j'} \mid V_k, V_l) = \lim_{\substack{\delta_j \to 0^+ \\ \delta_j \to 0^+}} \frac{\Lambda}{\Lambda + \Delta^{-j,j'} + \delta_j m_j + \delta_{j'} m_{j'}} = \frac{\Lambda}{\Lambda + \Delta^{-j,j'}}$$

iii. $$\lim_{\substack{\delta_j \to \infty \\ \delta_j \to \infty}} S^{flex-IJ,\lambda,\delta}(\delta_j, \delta_{j'} \mid V_k, V_l) = \lim_{\substack{\delta_j \to \infty \\ \delta_j \to \infty}} \frac{\Lambda}{\Lambda + \Delta^{-j,j'} + \delta_j m_j + \delta_{j'} m_{j'}} = 0$$

❐

PROOF of THEOREM 5:

i. trivial

ii. $$\lim_{\substack{\lambda_i \to 0^+ \\ \delta_j \to 0^+}} S^{flex-IJ,\lambda,\delta}(\lambda_i, \delta_j \mid V_k, V_l) = \lim_{\substack{\lambda_i \to 0^+ \\ \delta_j \to 0^+}} \frac{\Lambda^{-i} + \lambda_i m_i}{\Lambda^{-i} + \Delta^{-j} + \lambda_i m_i + \delta_j m_j} = \frac{\Lambda^{-i}}{\Lambda^{-i} + \Delta^{-j}}$$

iii. $$\lim_{\substack{\lambda_i \to \infty \\ \delta_j \to \infty}} S^{flex-IJ,\lambda,\delta}(\lambda_i, \delta_j \mid V_k, V_l) = \lim_{\substack{\lambda_i \to \infty \\ \delta_j \to \infty}} \frac{\Lambda^{-i} + \lambda_i m_i}{\Lambda^{-i} + \Delta^{-j} + \lambda_i m_i + \delta_j m_j}$$

$$= \lim_{\substack{\lambda_i \to \infty \\ \delta_j \to \infty}} \frac{c_1 + \lambda_i m_i}{c_2 + \lambda_i m_i + \delta_j m_j}$$

with c_1 and c_2 being positive constants. Then it is obvious that

$$0 \le \lim_{\substack{\lambda_i \to \infty \\ \delta_j \to \infty}} \frac{c_1 + \lambda_i m_i}{c_2 + \lambda_i m_i + \delta_j m_j} \le 1 \text{ and so}$$

$$0 = \lim_{\delta_j \to \infty} S^{flex-IJ,\lambda,\delta}(\delta_j \mid V_k, V_l) \le \lim_{\substack{\lambda_i \to \infty \\ \delta_j \to \infty}} S^{flex-IJ,\lambda,\delta}(\lambda_i, \delta_j \mid V_k, V_l)$$

$$\le \lim_{\lambda_i \to \infty} S^{flex-IJ,\lambda,\delta}(\lambda_i \mid V_k, V_l) = 1$$

❐