

## COEFFICIENTS OF DISSIMILARITY AND SIMILARITY WITH APPLICATIONS

**Idzi Siatkowski, Teresa Goszczurna, Alicja Szabelska,  
Joanna Zyprych**

Department of Mathematical and Statistical Methods  
Poznan University of Life Sciences  
Wojska Polskiego 28, 60-637 Poznań, Poland  
idzi@up.poznan.pl, teresag@up.poznan.pl,  
aszab@up.poznan.pl, zjoanna@up.poznan.pl

### Summary

Coefficients of dissimilarity: Euclidean, Rogers', Modified Rogers', Cavalli-Edwards', Reynolds', Nei's (1972), Nei's (1983) and coefficients of similarity: simple matching, Jaccard's, Dice's, Rogers and Tanimoto, Russel and Rao, Hamann, Ochiai are presented in this paper. Next, the possibility of their applications and calculation in R are considered. At the end, some example is presented.

**Key words and phrases:** coefficients of dissimilarity and similarity, genetics, R software

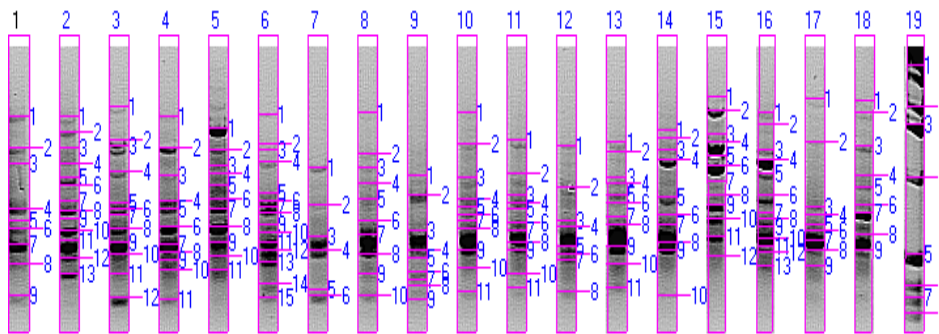
**Classification AMS 2010:** 62-07, 62-09, 62H20

### 1. Introduction

In the genetic research several coefficients of similarity and dissimilarity are introduced. They can help finding the genetic resemblance as well as genetic diversity between some populations or breeds, i.e. operational taxonomic units (OTU). The coefficients used in different kinds of investigations differ

from each other due to their mathematical properties. At first the coefficients of dissimilarity for allelic informative marker data will be introduced. As the next step the coefficients of similarity for allelic noninformative marker data will be presented.

First genetic example in this paper is based on five varieties of white clover: Romena (path 1 and L1) with 5 clones (path 2-6 and L2-L6), Rawo (path 7 and L7) with 2 clones (path 8-9 and L8-L9), Aura (path 10 and L10) with 4 clones (path 11-14 and L11-L14), CYMA (path 15 and L15) with 2 clones (path 16-17 and L16-L17) and Astra (path 18 and L18). DNA-based techniques like Random Amplified Polymorphic DNA (RAPD) were used for analysing the genetic diversity and variety of individual clones. Genomic DNA was extracted using the Thompson and Henry (1995) method. 12,5µl of polymerase chain reaction (PCR) mixtures contained: 25ng template DNA, 5U Tag DNA polymerase, Tag buffer (1M Tris/HCl, pH 8,3; and 25 mM KC), MgCl<sub>2</sub>; BSA; 2mM dNTP, and primer - 5 pmol/µl. Amplification were performed in a DNA thermocycler (T3 BIOMETRA). Amplification products were analyzed with the use of electrophoresis on 1.5% agarose gels in 1xTBE buffer, stained with ethidium bromide. 80 oligonucleotide primers were tested. The most polymorphic products were generated by primer OPA-04.



**Fig. 1.** RAPD banding pattern amplified by primer OPB-04, where path 19 – standard 1 KB DNA Ladder

On the basis of the generated products a table of molecular masses was drawn up. These results will be used for determination of the genetic similarities between the analyzed objects.

**Table 1.** Table of molecular masses for each variety and its clones

		MwRF																		
		L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13	L14	L15	L16	L17	L18	L19
11		1.805	1.805	2.036	1.805	1.587	1.850	1.120	1.896	1.038	1.896	1.440	1.368	1.416	1.562	2.275	1.896	2.227	2.179	3.054
12		1.345	1.538	1.440	1.345	1.321	1.392	0.803	1.275	0.878	1.392	1.058	0.940	1.298	1.465	1.942	1.636	1.416	1.896	2.036
13		1.163	1.345	1.368	1.038	1.141	1.321	0.592	1.120	0.627	1.018	0.878	0.690	1.018	1.321	1.513	1.321	0.788	1.368	1.636
14		0.774	1.163	1.078	0.832	1.058	1.185	0.523	0.971	0.523	0.847	0.817	0.652	0.971	1.207	1.416	1.207	0.731	1.038	1.018
15		0.690	1.018	0.817	0.774	0.955	0.893	0.378	0.847	0.494	0.788	0.774	0.541	0.878	0.863	1.252	1.058	0.664	0.940	0.506
16		0.639	0.955	0.774	0.652	0.847	0.832	0.352	0.690	0.450	0.731	0.677	0.506	0.817	0.731	1.141	0.878	0.627	0.803	0.396
17		0.551	0.832	0.731	0.551	0.774	0.788		0.604	0.421	0.690	0.627	0.488	0.717	0.664	0.987	0.745	0.571	0.745	0.344
18		0.476	0.759	0.639	0.514	0.652	0.745		0.506	0.396	0.639	0.561	0.369	0.664	0.561	0.878	0.704	0.523	0.604	0.298
19		0.352	0.704	0.561	0.488	0.561	0.664		0.443	0.337	0.523	0.523		0.541	0.523	0.803	0.652	0.469	0.523	
20			0.627	0.506	0.456	0.500	0.615		0.352		0.463	0.443		0.476	0.352	0.704	0.582			
21			0.571	0.443	0.337	0.456	0.561				0.369	0.387		0.387		0.592	0.541			
22			0.494	0.344			0.523									0.500	0.514			
23			0.443				0.482										0.469			
24							0.405													
25							0.344													

In the last section we will perform the analysis of this example using three of the coefficients presented below.

If the molecular marker data are allelic informative the coefficients can be calculated from the difference in the allele frequencies. In the case of allelic noninformative molecular marker data coefficients can be calculated based on absence or presence of observation of bands or signals. Thus in the paper we use the following notation:

$p_{ij}, q_{ij}$  – the frequencies of the  $j$ th allele at the  $i$ th locus in the two OTUs under consideration,

$n_i$  – the number of alleles at the  $i$ th locus,

$m$  – the number of loci,

$v_{ij}$  – the number of bands in common between both OTUs,

$w_{ij}$  – the number of bands present in the  $i$ th OTU and absent in the  $j$ th OTU,

$x_{ij}$  – the number of bands absent in the  $i$ th OTU and present in the  $j$ th OTU,

$y_{ij}$  – the number of bands absent from both OTUs.

Using this notation and data from Table 1 we can see for columns 1 and 2 (path 1 and 2) that  $v_{ij} = 3$  (3 bands have the same molecular masses),  $w_{ij} = 6$  and  $x_{ij} = 10$ .

## 2. Dissimilarity coefficients for allelic informative marker data

The first dissimilarity coefficient considered in this paper is **Euclidean** distance. It is defined as:

$$d_E = \sqrt{\sum_{i=1}^m \sum_{j=1}^{n_i} (p_{ij} - q_{ij})^2}.$$

The range of this measure is from 0 to  $\sqrt{2m}$ . It is a distance measure. It also fulfills the Euclidean property. However, the results can not be compared directly, since they depend on the number of loci.

The next coefficient is **Rogers'** distance (Rogers, 1972). It has the form:

$$d_{RO} = \frac{1}{m} \sum_{i=1}^m \sqrt{\frac{1}{2} \sum_{j=1}^{n_i} (p_{ij} - q_{ij})^2}.$$

The range of this coefficient is from 0 to 1. It is a distance. However, it has not the Euclidean property. It can be used in investigation of the assembly and validation of core collections and in determination of the pedigree relationships among operational taxonomic units.

The following coefficient is **Modified Rogers'** distance (Wright, 1978). It is defined as:

$$d_W = \frac{1}{\sqrt{2m}} \sqrt{\sum_{i=1}^m \sum_{j=1}^{n_i} (p_{ij} - q_{ij})^2}.$$

The range of this coefficient is from 0 to 1. It is a distance and it has the Euclidean property as well. It can be applied for prediction of heterosis with genetic dissimilarities and for the establishment of heterotic groups.

The **Cavalli-Sforza and Edwards' Chord** distance (Cavalli-Sforza and Edwards, 1967) is the next measure. It has the form:

$$d_{CE} = \sqrt{\frac{1}{m} \sum_{i=1}^m \left( 1 - \sum_{j=1}^{n_i} \sqrt{p_{ij} q_{ij}} \right)}.$$

The range of this coefficient is from 0 to 1. Similar to Modified Rogers' it is a distance with Euclidean property. Under assumption of selective drift model, it can be used for investigation of phylogenetic relationships among populations.

The subsequent coefficient is **Reynolds'** dissimilarity (Reynolds et al., 1983). It is defined as follows:

$$d_{RE} = -\ln(1 - \theta)$$

$$\text{where } \theta = \frac{\sum_{i=1}^m \left( \frac{1}{2} \sum_{j=1}^{n_i} (p_{ij} - q_{ij})^2 - \frac{1}{2n_i - 1} + \frac{1}{4n_i - 2} \sum_{j=1}^{n_i} (p_{ij}^2 + q_{ij}^2) \right)}{\sum_{i=1}^m \left( 1 - \sum_{j=1}^{n_i} p_{ij} q_{ij} \right)}.$$

The range of this coefficient is from 0 to  $\infty$ . It is neither a distance nor the Euclidean. Under assumption of selective drift model, investigation of phylogenetic relationships among populations can be performed.

The following coefficient is **Nei's** dissimilarity (Nei, 1972). It is calculated with the expression:

$$d_{N72} = -\ln \left( \frac{\sum_{i=1}^m \sum_{j=1}^{n_i} p_{ij} q_{ij}}{\sqrt{\sum_{i=1}^m \sum_{j=1}^{n_i} p_{ij}^2 \sum_{i=1}^m \sum_{j=1}^{n_i} q_{ij}^2}} \right).$$

The range of this factor is from 0 to  $\infty$ . It is neither a distance nor the Euclidean. It is determined on the base of mutation and drift - infinite-allele model. Under the assumptions of maintaining mutation-drift balance, the absence of selection and not large dissimilarity we have  $d_{N72} = 2\nu$ , where  $\nu$  is the muta-

tion rate per locus and generation,  $t$  is the time measured in generations after divergence of the two populations.

The second coefficient of Nei is the **Nei et al.'s** dissimilarity (Nei et al. 1983). It has the form:

$$d_{N83} = \frac{1}{m} \sum_{i=1}^m \left( 1 - \sum_{j=1}^{n_i} \sqrt{p_{ij}q_{ij}} \right).$$

The range of this factor is from 0 to 1. It is neither a distance nor the Euclidean. In the case of the homozygous inbred lines  $d_{N83} = d_{CE}^2$ .

### 3. Similarity coefficients for allelic noninformative marker data

The first coefficient of similarity is the **simple matching coefficient** (Sneath and Sokal, 1973). It has the form:

$$s_{SM} = \frac{v_{ij} + y_{ij}}{v_{ij} + w_{ij} + x_{ij} + y_{ij}}.$$

For this coefficient the dissimilarity coefficient  $d_{SM} = 1 - s_{SM}$  has a range from 0 to 1. It is a distance. However, it does not have the Euclidean property. For homozygous inbred lines  $d_{SM} = 1 - s_{SM} = d_R$ .

The second coefficient is the **Jaccard's** coefficient (Jaccard, 1908). It is defined as follows:

$$s_J = \frac{v_{ij}}{v_{ij} + w_{ij} + x_{ij}}.$$

Similarly to simple matching coefficient, the dissimilarity coefficient  $d_J = 1 - s_J$  has a range from 0 to 1. It is a distance without the Euclidean property.

The next coefficient is the **Dice's** coefficient (Dice, 1945). It has the form:

$$s_D = \frac{2v_{ij}}{2v_{ij} + w_{ij} + x_{ij}}.$$

The dissimilarity coefficient  $d_D = 1 - s_D$  has a range from 0 to 1. It is also called the Nei-Li distance. It is neither a distance nor Euclidean.

Beside the main three coefficients introduced above, there are several other coefficients of similarity that are used in genetic researches. Four of them are listed below:

- **Rogers and Tanimoto** coefficient (Rogers and Tanimoto, 1960):

$$s_{RT} = \frac{v_{ij} + y_{ij}}{v_{ij} + 2(w_{ij} + x_{ij}) + y_{ij}}.$$

- **Russel and Rao** coefficient (Russel and Rao, 1940):

$$s_{RR} = \frac{v_{ij}}{v_{ij} + w_{ij} + x_{ij} + y_{ij}}.$$

- **Hamann** coefficient (Hamann, 1961).

$$s_H = \frac{(v_{ij} + y_{ij}) - (w_{ij} + x_{ij})}{v_{ij} + w_{ij} + x_{ij} + y_{ij}}.$$

- **Ochiai** coefficient (Ochiai, 1957):

$$s_O = \frac{v_{ij}}{\sqrt{(v_{ij} + w_{ij})(v_{ij} + x_{ij})}}.$$

More details of applications of the coefficients of dissimilarity and similarity can be found in Balestre et al. (2009), Reif et al. (2005) and on <http://cran.r-project.org/web/packages/vegan/index.html>.

#### 4. Calculations in R

R software contains the functions for calculating genetic distances between populations as well as functions for computing coefficients of dissimilarity and similarity. Names of these functions with short descriptions are presented below. In parentheses names of packages are included.

-----  
**dist(stats)**

Description

This function computes and returns the distance matrix calculated by using the specified distance measure to compute the distances between the rows of a data matrix of **x**.

Usage

`dist(x, method = "euclidean", diag = FALSE, upper = FALSE, p = 2)`  
-----

**daisy(cluster)**

Description

Compute all the pairwise dissimilarities (distances) between observations in the data set of **x**.

Usage

`daisy(x, metric = c("euclidean", "manhattan", "gower"), stand = FALSE, type = list())`  
-----

**dist.genet(ade4)**

Description

This program computes measures of genetic distance from a set of gene frequencies in different populations with several loci.

Usage

`dist.genet(x, method = 1, diag = FALSE, upper = FALSE)`  
-----

**dist.genpop(adegenet)**

Description

This function computes measures of genetic distances between populations using a *genpop* object. The *genpop* object is a matrix of alleles' counts where genotypes are in rows and alleles are in columns. Currently, five distances are available, some of which are Euclidian.

Usage

`dist.genpop(x, method = 1, diag = FALSE, upper = FALSE)`  
-----

**vegdist(vegan)**

Description

The function computes dissimilarity indices.



Usage

```
vegdist(x, method="jaccard ", binary=FALSE, diag=FALSE, upper=FALSE,
na.rm = FALSE,...)
```

## 5. Example

The example is based on data presented in the Introduction. The analysis was performed using a platform R. First a matrix of distances (Table 2) calculated pairwise among varieties of white clover was constructed using **dist** function from package **ade4** and Euclidean coefficients of dissimilarity.

```
# program
library(ade4)
DIST<-dist(t(OPA04), method="euclidean")
# Euclidean coefficients of dissimilarity
```

**Table 2.** Dissimilarities between varieties of white clover using dist function

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13	L14	L15	L16	L17
L2	1.1																
L3	0.75	0.55															
L4	0.28	0.95	0.66														
L5	0.78	0.56	0.63	0.61													
L6	0.92	0.27	0.42	0.8	0.48												
L7	1.8	2.6	2.5	1.8	2.0	2.4											
L8	0.39	0.89	0.55	0.3	0.58	0.71	2.0										
L9	1.4	2.2	2.1	1.4	1.7	2.0	0.3	1.6									
L10	0.47	0.8	0.54	0.27	0.55	0.67	2.0	0.36	1.6								
L11	0.76	1.2	1.1	0.59	0.66	1.0	1.2	0.74	0.98	0.69							
L12	1.1	1.9	1.7	1.1	1.4	1.7	0.63	1.3	0.52	1.3	0.64						
L13	0.8	0.82	0.87	0.61	0.31	0.75	1.7	0.7	1.4	0.61	0.46	1.1					
L14	0.79	0.69	0.65	0.71	0.46	0.6	2.2	0.62	1.8	0.74	0.9	1.5	0.59				
L15	1.9	0.88	1.1	1.7	1.4	1.1	3.8	1.7	3.2	1.5	2.1	2.9	1.7	1.5			
L16	1.1	0.23	0.5	0.91	0.63	0.33	2.8	0.83	2.3	0.78	1.2	2.0	0.89	0.67	0.84		
L17	0.76	1.4	0.99	0.67	1.2	1.2	2.1	0.75	1.8	0.62	1.1	1.4	1.2	1.3	2.0	1.3	
L18	1.1	0.79	0.64	1.1	1.1	0.85	3.0	0.99	2.5	0.94	1.6	2.2	1.3	1.0	1.0	0.61	1.2

Below, a matrix of distances was created using **vegdist** function from package **vegan** and Jaccard's coefficients of similarity (Table 3).

```
# program
library(vegan)
Jaccard<-vegdist(OPA04,method="jaccard", na.rm=TRUE)
```

```
Jaccard<-as.matrix(Jaccard)
# Jaccard's coefficients of similarity
```

**Table 3.** Similarities between varieties of white clover using Jaccard's algorithm

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13	L14	L15	L16	L17
L2	0.23																
L3	0.17	0.13															
L4	0.06	0.22	0.15														
L5	0.18	0.13	0.10	0.15													
L6	0.19	0.05	0.10	0.19	0.11												
L7	0.41	0.52	0.50	0.42	0.45	0.50											
L8	0.09	0.21	0.13	0.08	0.13	0.16	0.45										
L9	0.34	0.49	0.45	0.35	0.42	0.47	0.10	0.38									
L10	0.12	0.19	0.10	0.06	0.12	0.15	0.44	0.09	0.39								
L11	0.17	0.28	0.21	0.12	0.17	0.25	0.33	0.16	0.30	0.13							
L12	0.25	0.41	0.37	0.26	0.33	0.38	0.20	0.30	0.14	0.31	0.19						
L13	0.19	0.19	0.14	0.14	0.07	0.15	0.41	0.13	0.38	0.10	0.11	0.29					
L14	0.17	0.15	0.11	0.15	0.11	0.12	0.47	0.13	0.42	0.14	0.17	0.34	0.13				
L15	0.36	0.16	0.23	0.35	0.27	0.19	0.61	0.33	0.58	0.31	0.39	0.51	0.32	0.28			
L16	0.23	0.06	0.12	0.22	0.13	0.06	0.53	0.18	0.49	0.17	0.27	0.41	0.18	0.13	0.16		
L17	0.13	0.28	0.19	0.12	0.24	0.24	0.42	0.15	0.36	0.14	0.19	0.26	0.23	0.23	0.34	0.26	
L18	0.23	0.14	0.09	0.21	0.15	0.13	0.54	0.17	0.49	0.16	0.27	0.42	0.19	0.15	0.17	0.11	0.21

### Acknowledgments

The authors wish to thank the Secretary and an anonymous reviewer by constructive comments and suggestions on an earlier version of the manuscript.

### References

- Balestre M., Von Pinho RG., Souza J.C., Lima J.L. (2009). Comparison of maize similarity and dissimilarity genetic coefficients based on microsatellite markers. *Genetics and Molecular Research* 7(3), 695–705.
- Cavalli-Sforza L.L., Edwards A.W.F. (1967). Phylogenetic analysis: Models and estimation procedures. *Am. J. Hum. Genet.* 19, 233–257.
- Dice L.R. (1945). Measures of the amount of ecologic association between species. *Ecology* 26, 297–302.
- Hamann U. (1961). Merkmalbestand und Verwandtschaftsbeziehungen den Farinosae: Ein Betrag zum System der Monokotyledonen. *Willdenowia* 2, 639–768.
- Jaccard P. (1908). Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.* 44, 223–270.

- Nei M. (1972). Genetic distance between populations. *Am. Nat.* 106, 283–292.
- Nei M., Tajima F., Tatenos Y. (1983). Accuracy of estimated phylo-genetic trees from molecular data. II. Gene frequency data. *J. Mol. Genet. Evol.* 19, 153–170.
- Ochiai A. (1957). Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. *Bull. Jpn. Soc. Sci. Fish.* 22, 526–53.
- Reif J.C., Melchinger A. E., Frisch M. (2005). Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. *Crop Science Society of America* 45, 1–7.
- Reynolds J., Weir B.S., Cockerham C.C. (1983). Estimation of the coancestry coefficient: Basis for a short-term genetic distance. *Genetics* 105, 767–779.
- Rogers J.S. (1972). Measures of genetic similarity and genetic distance. *Studies in Genetics*, Univ. Texas Publ. 7213, 145–153.
- Rogers D.J., Tanimoto T.T. (1960). A computer program for classifying plants. *Science* 132, 1115–1118.
- Russel P.F., Rao T.R. (1940). On habitat and association of species of anopheline larvae in south-eastern Madras. *J. Malaria Inst. India* 3, 153–178.
- Sneath P.H.A., Sokal R.R. (1973). *Numerical taxonomy*. Freeman, San Francisco, CA.
- Thompson D., Henry R. (1995). Single step protocol for preparation of plant tissue for analysis by PCR. *Biotechniques* 19, 394–400.
- Wright S. (1978). *Evolution and genetics of populations*. The Univ. of Chicago Press.

## WSPÓŁCZYNNIKI PODOBIEŃSTWA I NIEPODOBIEŃSTWA W ZASTOSOWANIACH

### Streszczenie

W pracy zostały przedstawione współczynniki niepodobieństwa: Euklidesowy, Rogera, zmodyfikowany Rogera, Cavalliego–Edwarda, Reynolda, Neia (1972), Neia (1983) oraz współczynniki podobieństwa: simple matching, Jaccarda, Dica, Rogersa and Tanimoto, Russela i Rao, Hamanna, Ochiaia. Następnie omówiono możliwości ich zastosowania oraz możliwości wyznaczania ich wartości wykorzystując platformę obliczeniową R. Jako ostatni punkt pracy przedstawiono przykład wyznaczania wartości wybranych współczynników.

**Słowa kluczowe:** współczynniki podobieństwa i niepodobieństwa, genetyka, ekologia, platforma R

**Klasyfikacja AMS 2010:** 62-07, 62-09, 62H20