Theor Appl Genet (1987) 75:194–199

Cluster analysis as a method for evaluation of genetic similarity in specific host – parasite interaction (*Lactuca sativa – Bremia lactucae*)

A. Lebeda and T. Jendrulek

VHJ Sempra, Plant Breeding Station, CS-798 17 Smržice; VHJ Sempra, Research Institute of Vegetable Growing and Breeding, CS-772 36 Olomouc, Czechoslovakia

Received June 10, 1987; Accepted July 17, 1987
Communicated by G. Wenzel

Summary. In gene-for-gene systems, specificity of host-parasite interactions is most often estimated qualitatively using the symbols +, –, (i.e. susceptibility and/or resistance). In large sets (interaction patterns) it becomes impossible to analyze numerous data by mere comparison. This is overcome by application of cluster analysis. In our experiments the methods in question were used to estimate the data obtained in a study on interactions between more than 220 *Lactuca sativa* cultivars and 12 *Bremia lactucae* physiological races (isolates) of Czechoslovak origin. The matrix of similarity coefficients was analyzed by hierarchical clustering. Similarity and/or dissimilarity of host R-genotypes was graphically expressed using the method of two principal components. The results obtained are related to genetic constitution of race specific resistance of the host and the possibility of predicting effective resistance sources.

Key words: Host/parasite interaction – Specific Resistance – Genetic similarity – Cluster analysis

Introduction

Phenotypical expression of specificity of host-parasite interaction can be estimated by numerous qualitative and/or quantitative methods. Qualitative evaluation is often accomplished in systems operating on a gene-for-gene basis, the reaction being expressed by "yes" or "no". Assessment of host-parasite relationships is carried out by means of semi-quantitative and quantitative methods (Johnson and Taylor 1976).

Data from studies of these interactions has been interpreted mostly without deeper mathematical insight. It seems necessary, however, to apply mathematical methods even in phytopathology, as stressed by Priestley et al. (1984) who have shown that in large data sets (more than 10 cultivars × 10 isolates) it becomes impractical to simply compare the infection data.

The methods of multivariate analysis and cluster analysis may serve as an example of the effective exploitation of a mathematical approach both in fundamental and applied research of host-parasite interaction specificity. The general use of cluster analysis in gene-for-gene systems was demonstrated by Lebeda and Jendrulek (1986, 1987a, b).

The method of cluster analysis makes it possible to group not only the genotypes with identical and/or similar interaction patterns, but also to estimate genetic determination of resistance. This paper is a methodical contribution of various aspects of application of some cluster analysis methods and possibilities of interpreting the results obtained in the systems carrying race-specific resistance.

Materials and methods

The previously published qualitative data concerning interaction between *Lactuca sativa* cultivars and Czechoslovak isolates of *Bremia lactucae* (Lebeda 1984) was used to demonstrate the usefulness of cluster analysis. This paper deals with the reaction of the world lettuce assortment towards 12 *B. lactucae* isolates. The set was subdivided into 29 groups according to interaction patterns and/or presence of R-factors (genes). The results obtained were interpreted on the basis of gene-for-gene theory (Person 1959), and its genetical modification for *L. sativa – B. lactucae* (Crute and Johnson 1976). In our previous study (Lebeda 1984), the R-factors (genes) were postulated by means of phenotype analysis and by comparing interaction patterns of known and unknown R-genotypes. The

B. lactucae isolates under study did not contain all combinations of v-factors to allow for precise determination of all known R-factors and/or their combinations. To compensate, data from literature was used to interpret the results (Lebeda 1984). Analysis of the initial set of objects (interaction patterns) was accomplished by means of complete unmodified data as published in the paper by Lebeda (1984) (Table 1). Some changes were made in the description of new R-factors (R5/8, R15, R17) according to Crute (personal communication 1985), Hulbert and Michelmore (1985), and Lebeda (1986a).

From the mathematical viewpoint, the genetic analysis was based on expression of similarity and/or dissimilarity of objects and respective groups by hierarchical clustering techniques. The method of principal components was applied for variation analysis of the set. All calculations were carried out using a set of programs for multidimensional data analysis AVIDAT (Jendrulek, unpublished).

Hierarchical clustering methods may be subdivided into agglomerative methods, which proceed by a series of successive fusions of the N objects into clusters, and divisive methods, which separate the set of N objects successively into finer groupings. We shall be concerned only with agglomerative methods, which operate on an inter-objects proximity matrix (similarity, dissimilarity, or distance matrix) and initially each object will be considered to be a separate, single-member "cluster". The two objects having the highest similarity (or smallest dissimilarity or distance) are then grouped together, and proximities between each of the remaining objects and this two-member cluster calculated according to one of the methods (average-linkage, single-linkage, complete-linkage, centroid, median, group-average, Ward-Wishart's).

In our analyses Jaccard's, Sokal-Michener's, and Rogers-Tanimoto's coefficients of similarity and Euclidean distance (Duran and Odell 1974; Lukasová and Šarmanová 1985) were used to express similarity, dissimilarity and/or distance in respective groups of R-genotypes. These coefficients have been found to be most feasible for genetic analysis of *B. lactucae* – *L. sativa* relationships (Lebeda and Jendrulek 1987a, b). Numerous clustering methods and their combinations were used together to express similarity, dissimilarity, and/or distance in a basic set consisting of 29 groups of lettuce R-genotypes. The set was analysed by means of the following methods: single and complete-linkage, group-average, centroid, unweighted group (median) and Ward-Wishart's. Results of the analyses were used to find either "stable" method-independent clusters, or objects with dependency on method of clustering.

Analysis of variability of the initial set was made by means of the method of principal components. Principal components analysis is essentially intended for use with quantitative characters. Gower (1966) shows that it may also legitimately be employed on binary data, although interpretation may be difficult (Dunn and Everitt 1982; Kendall 1971, 1975). A possible method for overcoming these problems is described by Williamson (1978). The principal components scores were also used for making principal component scatter plots.

The aim of the analysis was to compare relationships between cluster formation and/or structure of the initial set of objects, and genetic similarity of respective groups (R-genotypes) of lettuce.

The set of interaction patterns was also analyzed from the opposite viewpoint, i.e. various reactions of *B. lactucae* isolates. Basic information on application of multidimensional analysis methods for estimation of genetic variability of pathogen virulence has been reported by Lebeda and Jendrulek (1986, 1987a, b). The analysis was accomplished by means of the same mathematical apparatus as stated in the first part of this paper.

Table 1. Reaction patterns of the groups of lettuce cultivars to *B. lactucae* races arranged in a binary system according to the data by Lebeda (1984). 1 = compatible; 0 = incompatible interaction

Group of cultivars	Postulated resistance factors	CS races of *B. lactucae*											
		1	2	3	4	5	6	7	8	9	10	11	12
1	R0	1	1	1	1	1	1	1	1	1	1	1	1
2	R1	0	1	0	0	0	0	0	1	1	0	1	0
3	R2, 4	1	1	1	1	1	1	1	1	1	1	1	1
4	R3(+4?)	0	1	0	0	1	0	0	0	1	1	1	0
5	R5/8	1	1	1	1	0	1	1	1	1	1	1	0
6	R6	1	0	1	1	0	0	0	1	1	0	0	0
7	R7	0	1	0	0	0	0	0	0	1	0	1	0
8	R8	1	1	1	1	0	0	0	1	1	1	1	0
9	R9	0	1	1	0	0	0	0	1	1	0	1	0
10	R5/8, 10	1	1	1	1	0	1	1	1	1	1	1	0
11	R11	0	0	0	1	1	0	0	1	1	0	1	1
12	R12	1	1	1	1	1	0	1	1	1	1	1	1
13	R15	0	1	0	0	0	0	0	1	1	0	1	0
14	R14	0	1	1	1	1	1	1	1	1	1	1	0
15	R1, 2, 7	0	1	0	0	0	0	0	0	1	1	1	0
16	R2, 3	0	1	0	0	1	0	0	0	1	1	1	0
17	R2, 7	0	1	0	0	0	0	0	0	1	1	1	0
18	R1, 7	0	1	0	0	0	0	0	0	1	1	1	0
19	R2, 11	0	0	0	1	1	0	0	1	1	0	1	1
20	R3, 7	0	1	0	0	0	0	0	0	1	1	1	0
21	R3, 11	0	0	0	0	1	0	0	0	1	0	1	0
22	R6, 8	1	0	1	1	0	0	0	1	1	0	0	0
23	R7, 8	0	1	0	0	0	0	0	0	1	0	1	0
24	R2, 3, 7	0	1	0	0	0	0	0	0	1	1	1	0
25	R3, 7, 8	0	1	0	0	0	0	0	0	1	1	1	0
26	R6, 7, 8	0	0	0	0	0	0	0	0	1	0	0	0
27	R1, 4, 6, 7	0	0	0	0	0	0	0	0	1	0	0	0
28	R3, 6, 7, 8	0	0	0	0	0	0	0	0	1	0	0	0
29	R17	0	0	0	0	0	0	0	0	0	0	0	0

Results

Estimation of similarity among interaction patterns of lettuce cultivars

It seems evident from the previously published data that the set under study shows great variability in reactions to *B. lactucae.* At the level of distance 0 (Fig. 1) there are groups and/or their clusters (1,3; 5,10; 2,13; 7,23; 4,16; 15–25; 6,22; 11,19; 26–28), which have identical interaction patterns to CS *B. lactucae* isolates. However, these groups do not always exhibit genetic identity in resistance constitution. Several clear-cut clusters can be found at the level of distance 5–10 (Fig. 1). The first large cluster is formed by groups 1–8 comprising the factors of resistance R0; R2,4; R12; R14; R5/8; R5/8 + 10. This cluster contains R-factors which are either slightly effective or completely ineffective against CS *B. lactucae* isolates. The cluster is

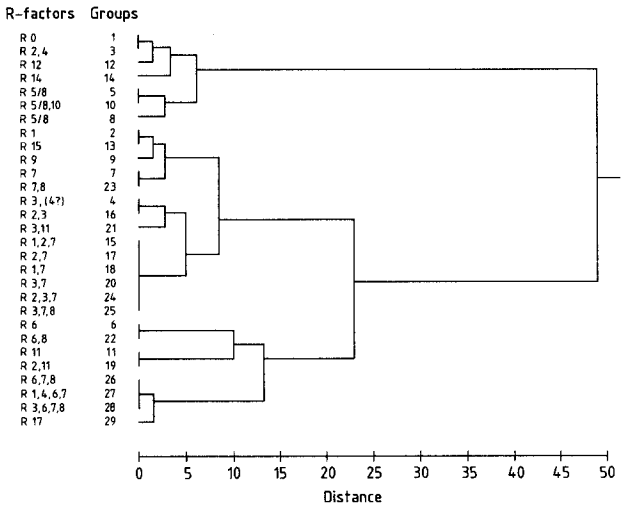

Fig. 1. A dendrogram (Ward-Wishart method) showing similarity and successive clustering of 29 groups of *L. sativa* cultivars

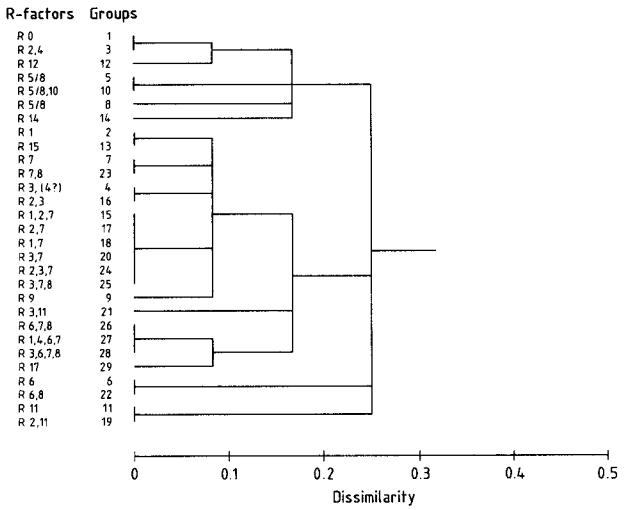

Fig. 2. A dendrogram (single linkage) showing similarity and successive clustering of 29 groups of *L. sativa* cultivars

conditioned by high frequency of complementary v-factors in the Czechoslovak population of the pathogen (Lebeda 1981, 1982). This cluster seems rather isolated; it fused with the next large group of clusters during the last step of clustering. It's structure was independent of the method of expressing dissimilarity and/or distance. The rest of the set under study represents a unique group comprising some typical clusters, especially 2–25 which contains groups 2–23 and 4–25. The former can be characterized by resistance to the isolates CS1, CS4, CS5, CS6, CS7, CS10, CS12, and susceptibility to CS2, CS9, CS11. Its formation is especially conditioned by the factors R1, R7, R9 and R15 which cannot be precisely differentiated by means of CS isolates. The latter (4–25) is typical by its absolute resistance to the isolates CS1, CS3, CS4, CS6, CS7, CS8, CS12, and consequently by the presence of factors R3, R7 and R11.

Group 21 (lines Type 83, Type 700) serves as an example of an unstable object whose classification is method-dependent. It is essential that this independent (single) object contributes to differences in the structure of clustering at higher levels as related to the groups 26–29 and 2–25. Its localization in the cluster 26–29 (Fig. 3) has resulted in greater independence of specific small clusters (6–22, 11–29), while its position according to Ward-Wishart's method (Fig. 1) breaks down this structure somewhat. Having examined all of the methods under consideration, the results shown in Fig. 3 can be regarded as the most instructive due to the structure of the studied set as related to groups 6–22 (determined by factor R6) and 1–19 (determined by factor R11). In spite of minute differences, the remarkable variation and independence of these clusters is a general phenomenon. The group 6–22 can be

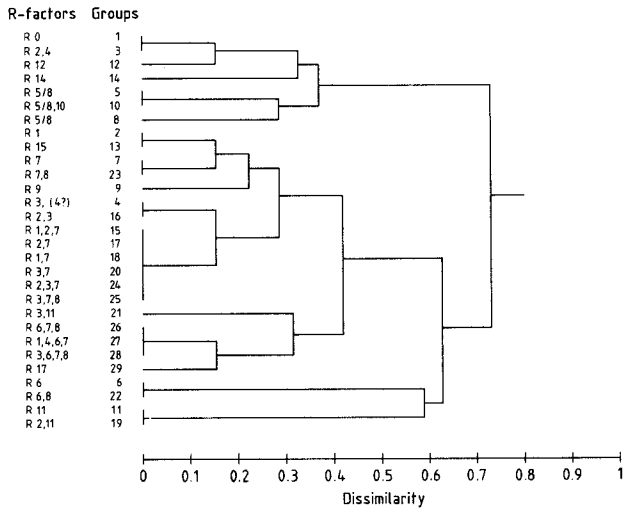

Fig. 3. A dendrogram (group average method) showing similarity and successive clustering of 29 groups of *L. sativa* cultivars

characterized by resistance to CS2 and CS11, (i.e. races with high frequency of susceptible reaction; Fig. 6). The group 11–19 differs from the others by resistance to CS2 and by susceptibility to CS12 whose reaction is, however, slightly virulent (Fig. 6). A typical feature of groups 26–28 is high complexity of a resistance genotype. An extreme example is the cultivar "Kinemontepas" (group 29) carrying absolute resistance to CS isolates.

The results obtained by means of Ward-Wishart's method as compared with the methods of single-linkage (Fig. 2) and group average (Fig. 3) differ to some extent in clustering of the respective groups. Similar to Ward-Wishart's method, the set of groups

(cultivars) can be divided into 3 large clusters [1–14, 2–29, 6–19 (method of single-linkage), 1–8, 2–29, 6–19 (method of group-average)]. It may be generally stated that most of the clusters are either similar or identical.

As seen in Fig. 1 the level of race-specific resistance to CS isolates is increasing in the direction from top to bottom (groups 1–29). R-factors of the clusters 1–8 are unsuitable for the conditions of Czechoslovakia and are not used in breeding. The largest group of clusters is 2–25 containing numerous genotypes and/or resistance genes (R3; R3,7; R3,11; R2,3,7); it is extensively exploited in breeding not only in Czechoslovakia but in many other European countries as well. The same applies for groups 6–29 (Fig. 1).

Relationships between genotype of resistance can be efficiently expressed by the method of principal components, which properly complements the data and information obtained from dendrograms. Figure 4 contains results of transformation of the basic set onto the plane of the first two principal components with indication of "stable" clusters obtained by means of cluster analysis (see "materials and methods"). In view of the fact that the first two principal components represent nearly 70% of the total variability of the set, the structure of clustering is in agreement with spatial representation of the respective groups of R-genotypes at the level of two principal components. The first principal component can be characterized by accumulation of virulence of the isolates CS3, CS4 and CS8. On the right-hand side of Fig. 4 groups of cultivars are characterized by susceptible reaction to the above-mentioned isolates. To the left of Fig. 4, the race-specific resistance of R-genotypes to the isolates in question increases. The second principle component is characterized by the variability factor conditioned especially by reaction to the isolates CS2, CS10 and CS11. It may be generally stated that R-genotypes resistant to the races under consideration are localized below (Fig. 4), while R-genotypes marked with dark points express the main typical groups of reactions and/or race-specific resistance to CS iolates.

R-genotype of group 29 represents absolute resistance to the isolates under study. A high level of field resistance is typical for the cluster of the groups 26–29 (Lebeda 1987; Norwood and Crute 1981). R-genotype of group 6 comprises the cultivars carrying resistance to isolates CS2, CS10–CS12 and susceptibility to CS1, CS3 and CS4. In the opposite corner of the Fig. 4 is group 15 which contains genotypes with converse reactions to the above-mentioned isolates (i.e. susceptibility to CS2, CS10–CS12 and resistance to CS1, CS3 and CS4). Reaction patterns of group 1 are characterized by absolute susceptibility to the isolates in question. Group 9 is positioned in the middle of Fig. 4 and is an example of a well-balanced reaction pattern.

Similarity of virulence among B. lactucae isolates

Variability of interaction patterns can be also evaluated from the pathogen viewpoint. The possibility of estimating genetic similarity of physiological races by means of their virulence phenotypes has been reported by Lebeda and Jendrulek (1987a). Similarity of virulence can be also judged on the basis of compatibility and/or incompatibility of the isolates in a set of host cultivars which need not be only differential.

Our data (Lebeda 1984) was evaluated by means of several methods of cluster analysis. In view of the fact that all the methods used gave similar results, we will only present a dendrogram obtained by Ward-Wishart's method (Fig. 5). The set of isolates under study was

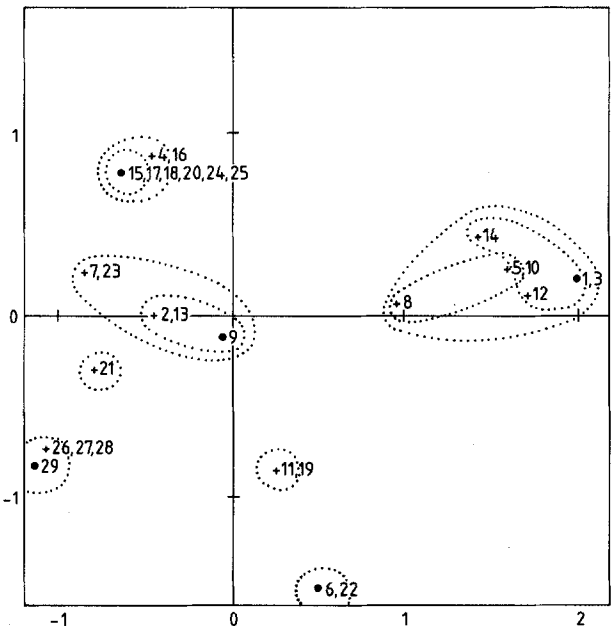

Fig. 4. Clusters of 29 groups of *L. sativa* cultivars as projected onto the plane of two principal components

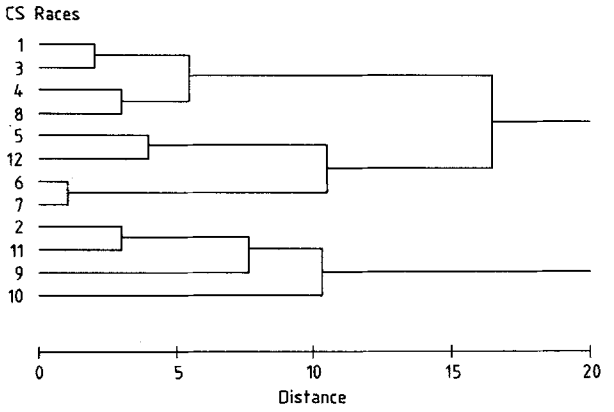

Fig. 5. A dendrogram (Ward-Wishart method) showing similarity and successive clustering of 12 *B. lactucae* races

divided into two to three large clusters. The first cluster is comprised of isolates CS1, CS3, CS4 and CS8, with medium to high level of virulence which can be characterized by considerable complexity of virulence phenotypes (in sensu Lebeda 1982) and absolute presence of the factors v2, v4, v5, v6, v8, and v10. The second, clear-cut group contains isolates CS2, CS9, CS10 and CS11, possessing high complexity of virulence phenotypes and absolute presence of the factors v2, v3, v4, v5, v7, and v10. The third cluster is formed by isolates CS6, CS7, CS5 and CS12 with extremely low complexity of the virulence phenotype and the lowest level of virulence as related to the set of cultivars under investigation.

Within the scope of large sets of host cultivars and virulence of pathogen, isolates can be judged not only by means of the presence of v-factors, but also by relative frequency of characters (i.e. taking into account the number of positive reactions of respective isolates as related to the total number of interactions under study, Fig. 6). This criterion was used to estimate virulence levels of 12 Czechoslovak *B. lactucae* races (Lebeda 1984). Fig. 6 shows that the race CS9 possesses the highest level of virulence as well as maximum frequency of positive reactions. The races CS2 and CS11 are also characterized by high levels of virulence. In most races, however, the level of virulence is from medium to low because relative frequency of the character barely reaches 50%.

Variability of the set of isolates can be also estimated by means of the involvement of the respective components (CS races) in the total variance (Fig. 7). In our set, the total variance reached more than 90% (Fig. 7), providing at least five components were involved (i.e. CS1 to CS5) which shows relatively high diversity of reaction patterns of the isolates.

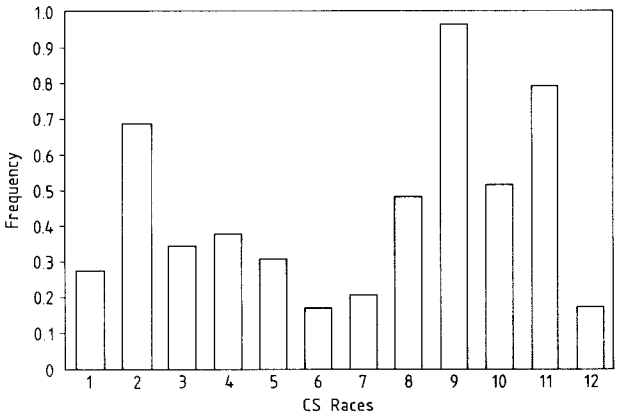

Fig. 6. Relative frequency of infection (set of cultivars) in a host-pathogen system of *L. sativa–B. lactucae*

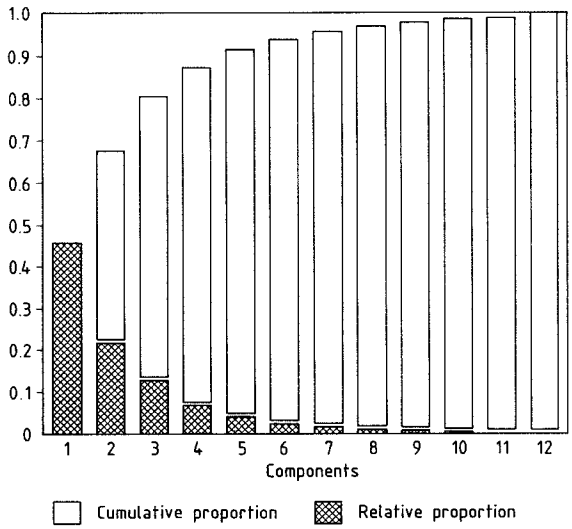

Fig. 7. A share of components (*B. lactucae* races) in the total variance

Discussion

In the gene-for-gene systems of host-parasite interactions (i.e. those based on race-specific resistance), phytopathologists most often express compatibility and/or incompatibility by means of binary coding (+, –; 1, 0) (Browder 1985; Loegering 1978). It is sometimes difficult to simply compare the results, especially under conditions of large variability which often occurs in obligate parasites (e.g. powdery and downy mildews, Crute 1985). This fact becomes extremely evident if there is no precise data available on the presence of R-factors (genes) and/or v-factors, and an approximate genetic analysis of the system cannot be carried out. In such a case, cluster analysis of qualitative data gives a good chance of exact estimation of host-parasite interactions as well as levels of similarity of resistance genotypes. The present example of the reaction of lettuce cultivars to *B. lactucae* has shown that the results obtained by means of cluster analysis cannot be simply identified with genetic constitution of resistance, especially in such cases when there is a lack of parasite isolates and/or virulence phenotypes available to distinquish among all R-factors. Despite these limitations, it seems evident that cluster analysis of qualitative data of interaction patterns may substantially contribute to exact detection of potential resistance sources and their approximate genetic similarity and/or dissimilarity. This advantage has been reported by Priestley et al. (1984) when comparing interactions between *Puccinia striiformis* isolates and winter wheat. Analysis of resistance variability compared with cluster analysis of virulences of parasite isolates (Lebeda and Jendrulek

1987 a, b) may contribute to improved prediction of effective host R-genotypes.

In our set under study, the application of various clustering methods (single-linkage, complete-linkage, centroid method, unweighted-group and average-group method, Ward-Wishart's method) gave nearly identical results differing from one another only in mere details. These methods exhibit different levels of sensitivity which have no substantial effect on the final interpretation, providing the structure of the initial data is good.

Our future work will be focused on verification of numerous association (similarity) coefficients reported in references (Hubálek 1982), which often either do not correspond with one another or provide heterogenous information.

The method of principle components potentiates to express similarity of resistance and/or susceptibility. Our results indicate that application of the method under consideration assures precise detection of R-genotypes and/or their groups as potential sources of resistance. The method is supposed to give reliable information on localization of race-specific resistance sources, and is comparable with the illustrative demonstration by means of minimum spanning tree as recommended by Priestley et al. (1984).

It may be concluded that methods of multidimensional analysis (e.g. cluster analysis) could substantially improve processing and correct interpretation of qualitative data in interaction patterns. The aim of our future work will be verification of such methods when applied to quantitative data estimation, which seems rather difficult by means of current statistical methods (Lebeda 1986 b, Scott et al. 1978).

Acknowledgement. We are indebted to Miss J. Potomková for translating the paper into English.

References

Browder LE (1985) Parasite : host : environment specificity in the cereal rusts. Annu Rev Phytopathol 23:201–222

Crute IR (1985) The genetic bases of relationships between microbial parasites and their hosts. In: Fraser RSS (ed) Mechanisms of resistance to plant diseases. Nijhoff/W Junk, Dordrecht, pp 80–142

Crute IR, Johnson AG (1976) The genetic relationship between races of *Bremia lactucae* and cultivars of *Lactuca sativa*. Ann Appl Biol 83:125–137

Dunn G, Everitt BS (1982) An introduction to mathematical taxonomy. Cambridge University Press, Cambridge

Duran BS, Odell PL (1974) Cluster analysis. Springer, Berlin

Gower JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53:325–338

Hubálek Z (1982) Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. Biol Rev 57:669–689

Hulbert SH, Michelmore RW (1985) Linkage analysis of genes for resistance to downy mildew (*Bremia lactucae*) in lettuce (*Lactuca sativa*). Theor Appl Genet 70:520–528

Johnson R, Taylor AJ (1976) Spore yield of pathogens in investigations of the race-specificity of host resistance. Annu Rev Phytopathol 14:97–119

Kendall DG (1971) Seriation from abundance matrics. In: Hodson FR, Kendall DG, Tautu P (eds) Mathematics in the archeological and historical sciences. Edinburgh University Press, Edinburgh, pp 215–252

Kendall DG (1975) Review lecture: The recovery of structure from fragmentary information. Philos Trans R Soc, Ser A 279:547–582

Lebeda A (1981) Population genetics of lettuce downy mildew (*Bremia lactucae*). Phytopathol Z 101:228–239

Lebeda A (1982) Geographic distribution of virulence factors in the Czechoslovakian population of *Bremia lactucae* Regel. Acta Phytopathol Acad Sci Hung 17:65–79

Lebeda A (1984) Race-specific factors of resistance to *Bremia lactucae* in the world assortment of lettuce. Sci Hortic 22: 23–32

Lebeda A (1986 a) *Bremia lactucae*. In: Lebeda A (ed) Methods of testing vegetable crops for resistance to plant pathogens. VHJ Sempra, Research Institute of Vegetable Growing and Breeding, Olomouc, pp 105–113

Lebeda A (1986 b) Specificity of interactions between wild *Lactuca* spp. and *Bremia lactucae* isolates from *Lactuca serriola*. J Phytopathol 117:54–64

Lebeda A (1987) Screening of lettuce cultivars for field resistance to downy mildew. Tests Agrochem Cult No 8 (Ann Appl Biol 110, Suppl), pp 146–147

Lebeda A, Jendrulek T (1986) Possibilities of host-pathogen relationship study by factor and cluster analysis. Proc 10th Czech Conf Plant Prot, Brno 2–5 September 1986, pp 13–14

Lebeda A, Jendrulek T (1987 a) Application of cluster analysis for establishment of genetic similarity in gene-for-gene host-parasite relationships. J Phytopathol 119:131–141

Lebeda A, Jendrulek T (1987 b) Methodological aspects of research on population genetic relationships between plants and plant pathogenic fungi. In: Lebeda A (ed) Application of phytopathological methods in vegetable breeding for resistance to diseases and pests. VHJ Sempra, Research Institute of Vegetable Growing and Breeding, Olomouc, pp 142–161

Loegering WQ (1978) Current concepts in interorganismal genetics. Annu Rev Phytopathol 16:309–320

Lukasová A, Šarmanová J (1985) Methods of cluster analysis. SNTL, Praha

Norwood JM, Crute IR (1981) Evaluation of lettuce cultivars for field resistance to downy mildew. Tests Agrochem Cult No 2 (Ann Appl Biol 97, Suppl), pp 74–75

Person C (1959) Gene-for-gene relationships in host-parasite systems. Can J Bot 37:1101–1130

Priestley RH, Bayles RA, Ryall J (1984) Identification of specific resistance against *Puccinia striiformis* (Yellow rust) in winter wheat varieties. II. Use of cluster analysis. J Nat Inst Agric Bot 16:477–485

Scott PR, Johnson R, Wolfe MS, Lowe HJ, Bennet FGA (1978) Host-specificity in cereal parasites in relation to their control. Plant Breeding Institute Cambridge, Annu Rep pp 27–62

Williamson MH (1978) The ordination of incidence data. J Ecol 66:911–920