Finding Unexpected Patterns in Microarray Data

Susana Perelman, María Agustina Mazzella, Jorge Muschietti, Tong Zhu, and Jorge J. Casal*

IFEVA, Facultad de Agronomía, Universidad de Buenos Aires, Av. San Martín 4453, 1417-Buenos Aires, Argentina (S.P., J.J.C.); Instituto de Investigaciones en Ingeniería Genética y Biología Molecular, Vuelta de Obligado 2490, 1428-Buenos Aires, Argentina (M.A.M., J.M.); and Syngenta Biotechnology, Inc., 3054 Cornwallis Road, Research Triangle Park, North Carolina 27709 (T.Z.)

We describe the performance of a protocol based on the sequential application of unsupervised and supervised methods to analyze microarray samples defined by a combination of factors. Correspondence analysis is used to visualize the emerging patterns of three set of novel or previously published data: photoreceptor mutants of Arabidopsis grown under different light/dark conditions, Arabidopsis exposed to different types of biotic and abiotic stress, and human acute leukemia. We find, for instance, that light has a dramatic effect on plants despite the absence of the four major photoreceptors, that bacterial-, fungal-, and viral-induced responses converge at later stages of attack, and that sample preparation procedures used in different hospitals have large effects on transcriptome patterns. We use canonical discriminant analysis to identify the genes associated with these patterns and hierarchical clustering to find groups of coregulated genes that are easily visualized in a second round of correspondence analysis and ordered tables. The unconventional combination of standard descriptive multivariate methods offers a previously unrecognized tool to uncover unexpected information.

Microarray data can be used for different purposes. In clinical studies, a common aim is to identify a set of genes whose expression signature is sufficiently specific for a given condition and can be used as markers in diagnosis. Biologists interested in understanding the mechanisms involved in a given process may want to find the genes whose expression is characteristic of a given set of experimental cases. There are basically three groups of multivariate approaches that can be followed to achieve this goal: class prediction methods, projection methods, and clustering. The first is defined as a "supervised" method whereas the other two are considered unsupervised because they do not reflect any previous knowledge or classification scheme (Slonim, 2002).

Class prediction methods such as support vector machines, decision tree algorithms, and discriminant analysis are specifically designed to classify objects into known groups and can be used to identify genes that characterize different conditions (Golub et al., 1999; Ramaswamy et al., 2001; Stephanopoulos et al., 2002).

Projection methods such as principal component analysis (PCA) or correspondence analysis (CA) can also be used to link cases to genes. These dimension reduction techniques allow the visualization of the structure of large microarray data sets in a few dimensions that retain a large amount of the original variation (Fellenberg et al., 2001; Misra et al., 2002). The matrix created by the genes (rows) and experimental conditions (columns) with expression values as entries, defines a multidimensional space where each gene represents a coordinate axis and the treatments are points located by their relative expression of these genes. Treatments with similar expression profiles occupy nearby positions in this space, and as more experiments are added to the data matrix, the number of points increases but the number of dimensions (genes) remains fixed. CA allows the representation of the treatments or cases in a low dimensional subspace (e.g. a plane of dimension two) revealing associations between them in expression profiles. Two treatment points close on the graph are likely to have a similar profile. The advantage of CA (Greenacre, 1984) over other dimension reduction techniques is that this projection can be achieved simultaneously for the two variables (genes and cases) yielding a “biplot” where genes and cases can be observed on the same plane, making the relationship visually more obvious (Fellenberg et al., 2001). The distance preserved through CA is the chi-square distance (ter Braak, 1985), which is equivalent to preserving the Euclidean distance (i.e. geometrical distance) between profiles of weighted conditional probabilities. This is a major difference with PCA, where the distance among objects is always, by definition, the Euclidean distance (Gower, 1982), because the relative positions of the objects in the rotated p-dimensional space of PCA are the same as in the p-dimensional space of the original descriptors. In general, CA may be applied to any data table that is dimensionally homogeneous (the physical dimensions of all the variables are the same, as is the case of...
gene expression in microarray data) and only contains non-negative values.

Clustering methods such as hierarchical clustering and k-means (Kaufman and Rousseeuw, 1990) can be used to organize both genes and cases into groups of roughly similar patterns. This procedure is useful to reveal close relationships but the broadest clusters defining the major patterns are difficult to visualize (Slonim, 2002).

When the number of experimental conditions is large, the wealth of information obtained from the samples hybridized to microarrays is useful not only to learn about the function of genes but also to learn about the samples themselves. The application of CA to a series of samples from synchronized cells of Brewer’s yeast (Saccharomyces cerevisiae) at different phases of the cell cycle suggested that some samples had been improperly classified because their position in the biplot was distant from that corresponding to other samples originally thought to correspond to the same phase (Fellenberg et al., 2001).

In the aforementioned example with Brewer’s yeast, the useful information about the samples could derive from the fact that 800 cell-cycle associated genes were used as starting point. Alternatively, CA could be intrinsically useful to uncover unknown relationships (visualized as convergence on the biplot) among experimental conditions, particularly when they result from the combination of two or more factors. To investigate this possibility, we applied CA to three sets of data. First, our own transcriptome data from seedlings of Arabidopsis involving the factorial combination of three light/dark conditions and four genotypes, the wild type (WT), the double mutant lacking the red and far-red light photoreceptors phytochrome A and B (phyA phyB), the double mutant lacking the blue-light photoreceptors cryptochrome 1 and 2 (cry1 cry2), and the quadruple mutant lacking phytochromes A and B and cryptochromes 1 and 2 (phyA phyB cry1 cry2). Second, previously published data from Arabidopsis plants exposed to different types of biotic (bacterial, fungal, and viral) or abiotic stress each one at different times after application (Chen et al., 2002). Third, the well-known leukemia data published by Golub et al. (1999), combining different subtypes of cancer, ages of patients, cell lineage immunophenotype, tissues, and institutional sources.

Supervised and unsupervised methods are normally used as alternatives, depending on the aims of the study. A second contribution of the work presented here is to demonstrate that these apparently contrasting approaches can be complementary. We describe a protocol combining a first unsupervised visualization of the whole data set through CA with the supervised exploration of the particular structures that emerged in this first step. This exploration is based on the discriminant loadings of the genes, derived from canonical discriminant analysis (CDA; Seber, 1984; Stephanopoulos et al., 2002) to identify the genes showing the best discrimination between experimental conditions or physiological states through their differential expression. CDA determines discriminant functions; i.e. some optimal combination of variables (genes), so that the first function provides the most overall discrimination between groups, the second provides second most, and so on. These functions are independent or orthogonal, i.e. their contributions to the discrimination between groups do not overlap. The relative weight of the variables in the functions is used to determine which variables best discriminate between two or more predetermined groups. Following the protocol proposed here, the reduced number of relevant genes that result from CDA is then grouped by using hierarchical clustering. The results are conveniently visualized by applying CA both to the genes identified by CDA and to the relevant samples. Finally, partial ordered tables of gene expression data provide new information that is easy to interpret. Implementation of the proposed protocol is simple because the multivariate methods involved (CA, CDA, and hierarchical clustering) are available in standard statistical packages. However, the application of the proposed protocol has consequences in terms of optimization of the design of microarray experiments to obtain more information.

RESULTS

Light Has Large Effects on Plant Transcriptome Even in the Absence of Four Major Photoreceptors

Light perceived by phytochromes A and B and cryptochromes 1 and 2 has profound effects on plant growth and development (Quail et al., 1995; Cashmore et al., 1999). The absence of these photoreceptors results in plants that retain the developmental pattern typical of darkness even when grown in the light (Fig. 1A; Mazzella et al., 2001). Dark-grown seedlings of the WT Arabidopsis and of the phytochrome A and phytochrome B double mutant (phyA phyB), cryptochrome 1 and cryptochrome 2 double mutant (cry1 cry2), and of the phyA phyB cry1 cry2 quadruple mutant were exposed to 1 or 3 h of white light or remained as dark controls before harvest. To find out the main patterns in gene expression, we applied CA (Greenacre, 1984), a descriptive tool used to find the best simultaneous representation in low dimension (number of axes) of the rows and columns of a data matrix (genes and light/genotype in our case). Because CA is affected by outliers, which forced the axes, we used rank transformation to expand the scale in the range of frequent values. The first two axes accounted for 53.3% of the variation and defined a plane where the different genotype-light combinations grouped mainly according to the dark, 1-h-light, or 3-h-light conditions, with less variation associated with the presence or absence of the four
major photoreceptors (Fig. 1B). This was surprising, given the paramount role played by these photoreceptors in plant development (Fig. 1A).

We used CDA to identify the genes that best discriminate among the three light/dark conditions. These genes were arranged in groups following common trends in their pattern of expression by using a hierarchical farthest-neighbor clustering algorithm, based on correlation coefficient as similarity measure. Instead of using the representation generated by CDA itself, the genes and the cases were displayed on the biplot generated by a second round of CA (Fig. 1C). This grouping was highly significant \((P < 0.0001)\) as indicated by a multivariate nonparametric test, multiresponse permutation procedure (MRPP; Mielke, 1984) applied to correlation measures. Most of these genes (Supplementary Table I) could be assigned a cellular and biochemical function. Light-induced genes included several genes involved in photosynthesis and electron transfer (e.g. cytochrome c), and genes involved in the phenylpropanoid biosynthesis pathway (e.g. flavonol synthase and chalcone synthase genes; Winkel-Shirley, 2001) probably involved in the production of secondary metabolites to protect from light damage. Among the light-repressed genes, the most conspicuous were those involved in cell wall loosening, necessary for cell growth, such as polygalacturonase (Hadfield and Bennet, 1998) and \(\beta\)-1,3-glucanase class I precursor (Cosgrove, 1999), and growth-hormone-regulated genes (e.g. putative auxin-induced protein and IAA20). The photoreceptor phototropin 1 mediates rapid and transient effects on growth (Folta and Spalding, 2001) and could be involved in transcriptome responses to light in the absence of phytochromes A and B and cryptochromes 1 and 2.

**Transcriptome Changes Induced by Bacterial, Viral, and Fungal Attack Converge at Later Stages in Arabidopsis**

We carried out CA with the complete matrix of data published by Chen et al. (2002) based on 402
potential stress-related genes encoding known or putative transcription factors (TF) monitored in various organs, at different developmental stages, and under various biotic and abiotic stresses. The most obvious conclusion is that the expression profiles of plants attacked by bacteria are very different from the patterns associated with other stresses (Fig. 2A), despite the fact that the bacterial attack group involves different pathogens. One of the advantages of the current approach is that the magnitude of the effect of bacterial attack can be easily placed into perspective relative to the variation caused by other conditions, including widely divergent developmental contexts for healthy plants (different ages, different organs, mock treatments used as controls for various stresses, and other ad-hoc controls). The plants exposed to abiotic stress and the samples from healthy plants (control and developmental treatment) formed two coherent groups. PCA applied to the same data set failed to show comparable patterns in the plane generated by the first two axes (Fig. 2B). One of the reasons of this failure is that the first axis ordered the samples according to the overall level of expression (data not shown) rather than according to the patterns of expression.

To gain precision in the search for responses in expression caused by the different types of stress, we left aside the cases involving healthy plants at different stages of development and calculated the deviations of each sample respect to its control profile. We then applied CDA to identify the TF genes that better discriminate between groups of stressors: abiotic, fungal, viral, and bacterial agents. The 146 TF genes selected by CDA were then clustered in five groups by hierarchical clustering (Supplementary Table II). CA was now applied to the resultant reduced matrix (146 genes × 33 treatments, rank-transformed data). Three homogeneous groups containing samples only from bacterial, viral, or abiotic stresses were obtained, but an additional group of cases still included fungal, viral, and bacterial attack (Fig. 3A). A closer inspection revealed that the samples of this apparently heterogeneous group of treatments had something in common as they corresponded to the latest time points used to characterize bacterial attack (30 h) or viral attack (5 and 14 d) in conjunction with the majority of fungal treatments. Thus, the ordination reveals that the TF expression profiles converge at late stages of attack by virus, bacteria, and fungi. If instead of using CA to visualize the cases, these are represented on the projection space generated by CDA itself, the convergence is no longer obvious (Fig. 3B). This reflects the fact that the CDA display maximizes discrimination whereas the cases are more freely ordered on the CA plot according to their extent of similarity.

To search for the TF genes that could express the apparent convergence in expression profiles associated with late biotic stress, we selected specific treatments narrowing the boundaries of the encompassed variability. The treatments included in this step of the analysis were: early virus attack (1 d, 5 cases), late virus attack (5 and 14 d, 6 cases), early bacterial attack (6 h, 9 cases), late bacterial attack (30 h, 2 cases), early fungal attack (12 h, 1 case), and late fungal attack (24–84 h, 5 cases); i.e. a total of 15 early stress cases and 13 late stress cases. We used CDA to identify the TF genes that best discriminate between early and late stress responses, i.e. only two groups (the analysis continued with the deviations from the respective controls). This procedure pointed to 30 TF genes that were arranged in two groups by hierar-

Figure 2. Placing Arabidopsis transcriptome changes into perspective. Samples corresponding to different developmental contexts (including controls of stress treatments; ●), plants infected with bacteria (▼), virus (★), or fungi (□), and plants exposed to abiotic stress (□) were analyzed by CA (A) or PCA (B). Data are from Chen et al. (2002).
A new round of CA was conducted with the reduced matrix (30 genes × 28 treatments), and the biplot is shown in Figure 3C. Although we selected the genes that best discriminate between early and late stress responses making no difference between stress type, bacterial and virus early responses still show clear differences in expression profiles. This grouping was highly significant (P < 0.0001) as indicated by MRPP. The first and third groups, dominated by AP2 domain and MYB-related TF, included TF that were induced respectively early or late in response to attack by bacteria, virus, or fungi. The second group contained TF that were induced early only by viral attack, including two auxin-induced TF (IAA8 and IAA17/AXR3-1).

The ordered table shows the correspondence between treatments and of TF genes (Fig. 3D, left) or between the average for each treatments group (e.g. all the samples corresponding to early bacterial attack regardless of the bacterial agent) and genes (Fig. 3D, right). Although the genes identified by CDA yield three clearly distinct groups of cases in the CA biplot (Fig. 3C), it is very difficult to find in the table individual genes with homogeneous response accounting for the differences among these groups of treatments (Fig. 3D, left), even if gene data are averaged for all bacteria, virus, or fungus attack samples (Fig. 3D, right). This is a strong evidence of the inherent multivariable nature of the gene expression responses, which determines the intrinsic caveat of hierarchical clustering.

Figure 3. Convergence of transcriptome changes induced in Arabidopsis either by infection with bacteria (bac), virus (vir), fungi (fun), or oomicetes (oo) or by abiotic stress (ab). A, CA applied to the deviations from control samples (calculated from TF data published by Chen et al. (2002); the numbers correspond to the sample number in the original publication). Ellipsoids were subjectively drawn to highlight groups of treatments. B, CDA projection of the same samples shown in A. C, Biplot (cases and genes) of the CA based on the genes (30 TF) that better discriminate between early and late attack by bacteria, fungi, and virus according to CDA. The genes are symbol coded according to hierarchical clustering. D, Ordered table for experimental conditions and TF genes shown in B. Right side of the table: treatment averages for the same data.
univariate (i.e. gene per gene) approaches to detect some underlying patterns of potential interest.

**Developmentally Biased Stress-Induced Genes in Arabidopsis**

An additional feature of the combined unsupervised/supervised protocol proposed here is that genes identified for their ability to discriminate among certain conditions can be used to visualize a different set of data in the search for unknown biologically meaningful relationships. For instance, we applied CA to the samples involving different developmental conditions by using the same 30 TF genes that best discriminate between early and late biotic stress (note that developmental samples were already out of the analysis at the point when these 30 genes were identified). The ordination of treatments on the plane generated by the first two axes of CA revealed three groups: one containing all root samples, another containing most leaf samples, and a third, more heterogeneous group containing seedling and reproductive tissue samples (Fig. 4). Most of the TF genes showing high expression in early bacterial stress are also highly expressed in leaves, whereas TF genes with high expression in late biotic stress showed relatively high expression in roots.

**Preparation Protocol Alters Transcriptome Patterns of Human Acute Leukemia Samples**

To investigate the applicability of the protocol presented here to a well-known data set, we used CA to characterize the patterns that emerges from 72 cases of patients with leukemia described by Golub et al. (1999). The samples differed in terms of tumor subtype (acute lymphoblastic leukemia [ALL] or acute myeloid leukemia [AML]), the age of the patients (adults or children), the tissue (bone marrow aspirates or peripheral blood), the cell lineage immunophenotype (T-lineage or B-lineage), and the institutional source of the samples. This set of data had been successfully analyzed by distinct class discovery procedures (Golub et al., 1999; Ramaswamy et al., 2001; Stephanopoulos et al., 2002) but failed to show gene expression patterns in a reduced space (Misra et al., 2002). The data were rank transformed before CA. The first two axes accounted for a small proportion of the variation (8% and 5%, respectively) but they allowed an obvious separation of ALL from AML cases (Fig. 5A). Axes 1 and 2 do not distinguish between age of the patients, type of tissue, cell lineage, or source of the sample. In other words, the various cases are well integrated into their corresponding ALL or AML subtypes. When the same set of data were analyzed by using PCA, these groups were not obvious (Fig. 5B; see also Misra et al., 2002). It is noteworthy that axes 3 and 4 derived from CA result in the segregation of the samples coming from St. Jude Children’s Hospital (Fig. 5C). These samples were prepared by using a very different protocol because they were subjected to hypotonic lysis rather than Ficoll sedimentation, and RNA was prepared by an aqueous extraction. Axes 3 and 4 also grouped differentially T-cells from B-cells (not shown in the figure). We used CDA to obtain the genes showing the best discrimination between St. Jude Children’s Hospital and other sources (Supplementary Table III). Hierarchical clustering yielded two groups, one with high and another with low expression in samples from St. Jude Children’s Hospital, and these genes are represented on the biplot produced by a second round of CA (Fig. 5D). The group showing low expression in St Jude Children’s Hospital samples contained genes associated with some types of cancer such as ETS2 V-ets avian erythroblasts virus E26 oncogene homolog 2 (Neznanov et al., 1999) and Tob (Tzachanis et al., 2001). The group with high expression contained some genes expressed in myeloid leukemia such as CGM6 Carcinoembryonic antigen gene family member 6 (NCA-95; Berling et al., 1999), CLC Charot-Leyden crystal protein gene (Paul et al., 1994), and three neutrophil elastase genes (Horwitz et al., 1999; Skold et al., 1999). It also included MPO myeloperoxidase, a specific bone marrow-expressed gene used in cytochemical studies of acute leukemia as marker to distinguished myeloid from lymphoid blast (Rouselet et al., 1995; Crisan et al., 1996). This grouping was highly significant ($p < 0.0001$) as indicated by MRPP.

---

**Figure 4.** Arabidopsis TF genes that discriminate between early and late biotic stress also group a different set of samples according to their developmental context. Ordination of development treatments in the plane generated by the first two axes of CA applied to expression profiles of 30 TF genes, which best discriminate between early and late biotic stress (Fig. 3B). Vegetative structures: L, leaf; R, root; St, stem; Se, seedling. Reproductive structures: Si, siliqua; F, flower; I, inflorescence. Ellipsoids were subjectively drawn to identify groups of conditions.
DISCUSSION

We describe a powerful protocol combining unsupervised and supervised methods to extract biologically meaningful information from microarray data (Fig. 6): CA is used to investigate the major underlying patterns of the cases from which the samples are extracted. CDA is then applied to identify the genes that cause the pattern of interest. Finally, a second round of CA based on these genes (grouped by using hierarchical clustering) is used to visualize both cases and genes (biplot). The strength and novelty of the approach is given by the combination of methods previously used only in isolation.

CA allows the representation of cases and genes on the same plane (biplot) and can therefore be used to find the genes whose expression correlates with a given set of samples (Fellenberg et al., 2001). Here, we show that CA can also be used to uncover unexpected features of the experimental cases themselves by showing that some of these cases are more and others less related to each other than previously thought. We have observed, for instance, that the expression of many genes is affected by the transition between darkness and light even in the absence of the four major plant photoreceptors. The clue for this pattern was provided by the observation that on the first axes of the CA, the cases grouped according to darkness or light exposure rather than according to the photoreceptor mutations. CA is particularly useful when microarray experiments incorporate a large number of treatments or conditions defined by combination of factors (e.g. light/dark conditions × photoreceptor mutations, type of stress × duration of the stress, etc.) because there is ample room for unexpected patterns. This feature suggests that complex experiments involving the combination of several factors simultaneously defining each treatment would be potentially more informative than many experiments with more simple treatments.

Figure 5. Visualization of human cancer data by CA groups different tumor sub-classes but also different hospital sources. ALL (●) and AML (□) data published by Golub et al. (1999) were analyzed by CA (A) or PCA (B), and the first two axes are shown in each case. C, Axes 3 and 4 of the same CA separate the samples coming from St. Jude Children’s Hospital (▲) from other sources (○). D, Biplot (cases and genes) of the CA based on the genes that better discriminate between from St. Jude Children’s Hospital (▲) and other tumor sources (○) according to CDA. The genes with higher (+) or lower (−) expression in samples from St Jude Children’s Hospital were grouped according to hierarchical clustering.
The comparative analysis revealed the better performance of CA over PCA for dimension reduction of microarray profiles. The first principal components of PCA are strongly affected by size (in this case the absolute values of expression of a microarray sample) because the method is based on Euclidean distances (Seber, 1984). This means that the Euclidean distances among objects are preserved through the rotation of axes, regardless of the use of rough data (covariance matrix) or standardized data (correlation matrix; Gower, 1966). In CA, the distance used to quantify the relationship among rows and columns is chi-square, which is low when the profiles of two vectors show similar shape, independent of their absolute value.

We confirm the usefulness of discriminant loadings derived from CDA to find the genes that best discriminate among classes. However, according to the protocol proposed here, these groups do not necessarily reflect previous knowledge on the subject because they can result from the unsupervised aggregation of experimental conditions resulting from CA. The combination between unsupervised and supervised methods is extended to the use of hierarchical clustering applied not to the original set of data but to the reduced set of genes emerging from CDA, thus eliminating the drawback of using hierarchical clustering with large bodies of data (Slonim, 2002).

Another unconventional feature of the protocol presented here is the use of a second round of CA based on the genes emerging from CDA to visualize both genes and cases on the biplot, rather than using the CDA projection itself. The CDA display maximizes the separation among pre-ordained sample classes. The display of CA is not subjected to this restriction; the samples are freely ordered according to their similarity and can reveal unexpected patterns. For instance, we used CDA to find the TF genes that better discriminate among different types of biotic and abiotic stress. The CA display confirms that the genes are able to discriminate among cases (Fig. 3B). The CA plot reveals that the genes that discriminate among types of biotic stress show convergence of their patterns at late stages of attack, a previously unsuspected pattern.

The enhanced visualization of the data also makes evident the multivariate nature of microarray data. Although the groups of samples are clearly distinguishable, it is difficult to find a single gene able by itself to discriminate among the different treatments. Rather it is the behavior of the groups of genes as a whole that provides discrimination among treatments. This underscores the need to use multivariate methods for the analysis of microarray data as significant patterns may be ignored in a gene-by-gene analysis.

Another application of the combination of supervised and unsupervised methods is the ability to use a set of genes found by CDA to discriminate among a given set of treatments to classify an unrelated set of cases. These cases are represented on a plane generated by CA based on the previously identified discriminating genes. Following this procedure, we have observed that genes that discriminate between early and late biotic attack are differentially expressed in different plant tissues.

Microarray data present two different, unrelated challenges. One is the efficient use of the potential of the technique because important biological information may remain buried in the large amount of data. The other is the statistical treatment of the results because, given the large number of genes and the generally small number of replicates, false positive genes are not unlikely. The protocol proposed here addresses only the first of these two issues, and the probabilities provided in our analysis indicate that the group patterns are unlikely to be found by mere chance. This does not imply that the pattern of expression of each of the genes of a given group is statistically significant. Testing the latter requires the application of a separate set of techniques and very often, additional experimental follow-up.

**MATERIALS AND METHODS**

**Sample Processing and Hybridizations**

The WT of Arabidopsis and all photoreceptor mutants were in the Landsberg erecta background. The production of multiple mutants has been described previously (Yanovsky et al., 2000; Mazzella et al., 2001). Seeds of each genotype were sown on 0.8% (w/v) agar in clear plastic boxes (40 × 33
Multivariate Analysis Protocol

The proposed protocol consists of a first round of CA (PROC CORRESPOND, SAS/STAT V8.02; for details of the technique, see Fellenberg et al., 2001; Greenacre, 1984) applied to the complete matrix of gene expression data rank-transformed within each sample. Selected cases (experimental conditions, genotypes, or tissues) were grouped according to the patterns emerging in the first axes of CA. To identify the genes that best discriminate among the different groups we used CDA (PROC CANDISC, SAS/STAT V6) applied to the reduced matrix of selected cases. Genes with extreme loading in the first k-1 discriminant coordinates were retained for further analysis (cutting level of absolute loading = 0.4). Selected genes were grouped by using farthest-neighbor clustering applied to correlation coefficients (CLUSTER, PC-ORD V4; McCune and Meflord, 1999). Finally, CA was applied to the reduced matrix, and information about cases and groups of genes was displayed on the plane of the resulting ordination to facilitate the biological interpretation of the results.

The hypothesis of no difference between groups of entities (either samples or genes) was tested by using a multivariate non-parametric procedure (MRPP, PC-ORD V4; Mielke, 1984). MRPP has the advantage of not requiring assumptions such as multivariate normality and homogeneity of vari- ances. The method calculates the average distance within each group and its weighted mean for the whole set of groups, where the weight depends on the number of items in each group. The P value reported here is the probability of finding weighted mean distances equal to or smaller than that observed when all possible partitioning of the same sizes are generated.

ACKNOWLEDGMENT

We thank Dr. Marcelo Yanovsky (IFEVA) for his valuable help with sample collection.

Received June 17, 2003; returned for revision July 23, 2003; accepted August 31, 2003.

LITERATURE CITED


Ter Braak CJF (1985) Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. Biometrics 41: 859–873


