Assessment of Hierarchical Clustering Methodologies for Proteomic Data Mining J. Proteome Res. 2007, 6 (1), 358−366.

ARTICLE *in* JOURNAL OF PROTEOME RESEARCH · FEBRUARY 2007

Impact Factor: 4.25 · DOI: 10.1021/pr078001e

6 AUTHORS, INCLUDING:

Emilie Dumas
Claude Bernard University Lyon 1
13 PUBLICATIONS **313** CITATIONS
SEE PROFILE

Isabelle Piec
University of East Anglia, Norwich, United Ki…
22 PUBLICATIONS **314** CITATIONS
SEE PROFILE

Daniel M Bechet
French National Institute for Agricultural Res…
78 PUBLICATIONS **4,095** CITATIONS
SEE PROFILE

Jean-François Hocquette
French National Institute for Agricultural Res…
276 PUBLICATIONS **4,409** CITATIONS
SEE PROFILE

technical notes Journal of proteome research

Assessment of Hierarchical Clustering Methodologies for Proteomic Data Mining

Bruno Meunier,*,† Emilie Dumas,‡ Isabelle Piec,§ Daniel Béchet,|| Michel Hébraud,‡,⊥ and Jean-François Hocquette†

UR 1213, Unité de Recherches sur les Herbivores, Equipe Croissance et Métabolisme du Muscle, INRA de Clermont-Ferrand/Theix, E-63122 Saint-Genès Champanelle, France, UR454 Microbiologie, Equipe Qualité et Sécurité des Aliments (QuaSA), INRA de Clermont-Ferrand/Theix, F-63122 Saint-Genès Champanelle, France, Laboratoire d'Immunologie EMI 0351, INSERM, 3 rue des Louvels, E-80036 Amiens, France, UMR 1019, Unité de Nutrition Humaine, INRA de Clermont-Ferrand/Theix, F-63122 Saint-Genès Champanelle, France, and Plate-forme protéomique, INRA de Clermont-Ferrand/Theix, E-63122 Saint-Genès Champanelle, France

Received July 12, 2006

Abstract: Hierarchical clustering methodology is a powerful data mining approach for a first exploration of proteomic data. It enables samples or proteins to be grouped blindly according to their expression profiles. Nevertheless, the clustering results depend on parameters such as data preprocessing, between-profile similarity measurement, and the dendrogram construction procedure. We assessed several clustering strategies by calculating the *F*-measure, a widely used quality metric. The combination, on logged matrix, of Pearson correlation and Ward's methods for data aggregation is among the best clustering strategies, at least with the data sets we studied. This study was carried out using PermutMatrix, a freely available software derived from transcriptomics.

Keywords: proteomics • bioinformatics • data mining • hierarchical clustering • 2-D PAGE

Introduction

Quantitative proteomics using two-dimensional polyacrylamide gel electrophoresis techniques (2-D PAGE) generates ever more data, building a growing haystack in which the biochemist hopes to find the miraculous needle. The main difficulty in exploring these data lies in the high number of variables, in this particular case, the relative abundance of hundreds or thousands of protein spots ($P \gg 100$), compared to the small number ($n_{bio} < 10$ or more) of individuals or biological samples of interest analyzed by 2-D PAGE with often few replications ($n_{exp} < 3-5$). Thus, data mining is an important tool for a first exploration of proteomic results, since the amount of data rises with the increasing sensitivity and availability of the technique, although the data quality remains insufficient from a statistical point of view. In fact, abundance measurement distribution on a single gel is often asymmetrical and extended over a wide dynamic range. In addition, the data matrix may have many missing values (MV) due to a real absence of protein spots or miss-detection by the image analysis software due to any technical difficulty, from protein extraction to electrophoresis. In conclusion, proteomic experiments may prove to be a pitfall if special care is not taken and notably during bioinformatic analysis.[1]

The first objective of proteomic data mining is differential expression analysis. This consists of comparing two or more predefined biological conditions to accurately highlight the few protein spots of interest among noise. Current statistical methods give good results when data are correctly preprocessed.[2,3]

In this paper, we were interested in the second objective of proteomic data mining, that is, the need to group or to classify individuals according to their global expression profiles, without any *a priori* knowledge of the biological reasons for the existence of these groups. This may be of great interest for understanding complex biological systems, such as characterizing tumor samples for clinical diagnostics,[4,5] or investigating the genetic variability of different ecotypes, for example.[6] Thus, Appel et al.[7] introduced early (1988) the automatic classification of 2-D PAGE in the MELANIE program. In the same way, it may also be interesting to group protein spots with similar expression profiles among the samples,[8] for example, in kinetic studies[9,10] or for discovering biomarkers of a specific biological function. To this end, several multivariate analysis methods have been reviewed.[11,12] Here, we focus on the hierarchical clustering analysis (HCA), mainly used in proteomics.

HCA is a technique of choice for exploring and visualizing large data sets, such as in transcriptomic studies, with the analysis of microarrays.[13] Briefly, HCA consists in calculating the dissimilarity, usually called the distance, between the individuals (each one being analyzed with one chip or in one gel) with one individual corresponding generally to one column of the data matrix. In the case of two-way HCA, the same procedure is carried out on the variables (the genes or the proteins) which correspond to the rows of the data matrix.

* To whom correspondence should be addressed. Phone: +33 473624097. Fax: +33 473624639. E-mail: bruno.meunier@clermont.inra.fr.

† UR1213, Unité de Recherches sur les Herbivores, Equipe Croissance et Métabolisme du Muscle, INRA de Clermont-Ferrand/Theix.

‡ UR454 Microbiologie, Equipe Qualité et Sécurité des Aliments (QuaSA), INRA de Clermont-Ferrand/Theix.

§ Laboratoire d'Immunologie EMI 0351, INSERM.

|| UMR1019, Unité de Nutrition Humaine, INRA de Clermont-Ferrand/Theix.

⊥ Plate-forme protéomique, INRA de Clermont-Ferrand/Theix.

10.1021/pr060343h CCC: $37.00
Published on Web 12/14/2006

About 87% of the 28 recent proteomic publications analyzed in this study preferred the Pearson correlation coefficient or the Euclidean distance for HCA (supplementary Table 1, Supporting Information). Once the distance matrix is calculated, an agglomerative clustering algorithm is generally performed to construct the final dendrogram, a tree-like representation in which the closest objects or clusters (connected with smallest branches) are mathematically the most similar. There too, several algorithms can be employed. In proteomics, the unweighted paired group average linkage (UPGMA), complete linkage, and Ward's methods are the most commonly used aggregation procedures. A synthetic description and comparison of these clustering methodologies applied to gene expression clustering is reviewed in D'haeseleer et al..[14] Therefore, HCA needs to be tuned, first for selection of the distance measurement and second for the aggregation procedure. At present, it appears difficult for the biochemist to make this choice on his own, since these parameters are not always clearly described in the proteomic literature. Moreover, the clustering result is sometimes announced as biologically doubtful in some of the publications, which leads the authors to declare that HCA is not relevant when applied to their data. Preprocessing of data is almost never mentioned, although this step is known to be essential. Finally, we identified as many statistical tools needed for HCA as publications, which also complicates the choice. Unlike proteomics, transcriptomics helps to cope with these difficulties, proposing a lot of free (or costly) bioinformatic tools and keys for their utilization. The effect of missing values, need for data transformation, and effect of clustering methods have also been largely evaluated.[15]

This article aims, therefore, to determine whether one of the tools and the methods developed for microarray technologies could be employed to improve and to facilitate HCA of complex proteomic data sets. PermutMatrix software[16] has been identified as one of the few freely available programs permitting evaluation of the overall major HCA methods described previously. Two real and complementary proteomic data sets have been used to evaluate the effects of distance and the aggregation procedure on the clustering results. Furthermore, several innovative data preprocessing operations are proposed to potentially enhance the quality of the results arising from HCA methods. All these methodologies are evaluated and then discussed in this paper.

Materials and Methods

Data Set Description. The first data set (DS1) results from a proteomic analysis which was performed on 12 *Listeria monocytogenes* strains, a food-borne pathogen bacterial species. The strains belonged to three different serovars (1/2a, 1/2b, and 4b) and had three different origins and levels of virulence. For each strain, two protein extractions were made from two independent cultures, and at least three 2-D PAGE gels with 50 μg of protein were performed per protein sample. Finally, more than 72 gels were performed. Each run for the 2-D PAGE separation was carried out with 12 (first dimension) or 6 (second dimension) protein samples obtained from 6 different bacterial strains which allowed the technical variability to be taken into account. Silver-stained 2-D PAGE gels were scanned using a GS-800 imaging densitometer (Bio-Rad Laboratories), and image analysis was performed using Image Master 2D Platinum (GE Healthcare). Saturated spots were excluded in order to work in the linear range of silver nitrate staining. The five most reproducible gels were selected for each strain by using the scatter plot analysis tool (experimental replications *per* condition n_{epx}= 5). The 60 selected gels for the 12 bacterial strains were then matched with a reference gel. The reference gel was produced with a mixture of equal quantities of protein extracts from the 12 strains to visualize a compilation of all the protein spots. Each gel (ex: S01_1 for the gel 1 of the strain 1) was then characterized by its expression profile, a vector of 599 relative volume (% vol) values (corresponding to the P = 599 protein spots matched across the N = 60 gels). The final DS1 consists of a [599 × 60] volume matrix (VM) with about 36% of missing values.

The second data set (DS2) concerns a differential proteomic analysis which investigated rat age-related sarcopenia.[17] This study was undertaken on 3 groups of 5 rats (samples A−F) slaughtered at the age of 7, 18, or 30 months. 2-D PAGE gels were made in triplicate (1, 2, 3) on each individual protein extract (biological replications *per* condition n_{bio} = 5, experimental replications *per* condition n_{epx} = 3). The batch composition was carefully randomized to take the experimental variability into account. The 45 colloidal blue-stained gel images were acquired using a GS-800 imaging densitometer, and image analysis was performed using PDQuest 2-D analysis software (Bio-Rad Laboratories). Each gel (ex: 7mA1 for gel 1 of rat A sacrificed at 7 months) was then characterized by its expression profile, a vector of 341 relative volume (% vol) values (corresponding to the P = 341 protein spots matched across the N = 45 gels). The final DS2 consists of a [341 × 45] volume matrix (VM) with about 4% of missing values.

Data Preprocessing Step 1: Missing Value Imputation. Experimental conditions largely affect protein expression, revealing strong disparities in the data files. The first is missing values, which are known to be very disturbing for clustering algorithms.[18,19] To deal with this problem and complete the data matrix, one solution is to convert the original matrix to a binary matrix (VM → VM-Bin). Existing values are thus replaced by "1" and missing values are replaced by "0". This first approach is well-accepted, but it reduces the richness of information drastically, since the quantitative measures are permanently lost.

A second solution consists in replacing all missing values by zero (VM → VM-Zero). This solution makes the hypothesis that a missing value corresponds, in most cases, to a protein undetectable by staining due to its very low expression level. This is not always true, which is why this approach is highly criticized in transcriptomics where the *k*-nearest neighbor (KNN) method is generally advised for imputing missing values.[19] But, in proteomics, the origin of missing values is fundamentally different, and the number of missing values may be too important (>10%) for an efficient imputation with a KNN algorithm. These last considerations are, in fact, valid for both procedures (VM-Bin, VM-Zero). Finally, the two processed data matrices (VM-Bin, VM-Zero) remain the same size [P × N] as the original one (VM), so that no protein spot is discarded before further analysis. This is the only advantage of these two procedures. Now we consider that these two MV imputation procedures are correct, since extreme care was taken over gel-making for both data sets. Consequently, we may suppose that missing values did not occur by random chance but were biologically meaningful.

A third solution is nevertheless considered. This consists in working only on the reliable protein spots. Thus, a protein spot is kept if at least (n_{epx} − 1) volume values are available for all samples. Then, any missing value in a sample is replaced by

the mean of the existing values for this protein spot in that sample. In the other cases, the protein spot is definitely excluded. This procedure results in a reduction of the original data matrix (VM) to a reliable one (VMR) of smaller size [$P' \times N$], $P' \leq P$. Even if this procedure is not very stringent, it is nevertheless useful when a reasonable number of replications is available ($n_{epx} > 3$).

Finally, it is important to note that, before HCA is applied, protein spots of interest (statistically significant ones, for example, or proteins with sufficient variance) do not have to be selected using an ANOVA filter as is usually done in proteomics.[20] We thus preferred here to deal with all the protein spots, without any *a-priori* knowledge or without any prior analysis (such as ANOVA), even if proteins whose variation is due only to experimental conditions may be disturbing for any clustering algorithm.

Data Preprocessing Step 2: Volume Normalization. A second problem inherent to proteomic data sets is the result of the wide dynamic range characteristic of the volume distribution. Thus, certain very abundant protein spots may weigh much more on the clustering results than the majority of weakly abundant ones. To deal with this problem, Vohradsky[8] proposed dividing the expression profiles of the samples by the maximum value. In the same way, Gion et al. proposed a log transformation.[21] We chose a logged ratio-based approach that also acts as a data normalization step. This type of procedure is classically used for transcriptomics because fold change is indeed an interesting way to investigate differential expression.[3] To this end, each protein spot volume was divided by the mean of all the existing values for this protein spot, present in all the N gels. Unfortunately, this procedure can only be performed on the reliable data matrix (VMR) because the ratio cannot be calculated on zero or missing values. This ratio-based data matrix was then submitted to the usual base 2 logarithmic transformation (VMR-Ratio). This approach allowed each sample (a single gel) to be compared to a common mean sample of reference (a synthetic average gel), so that all the samples could then be compared one-to-one (0, no difference with the mean; +1, 2-fold greater than the mean; −1, 2-fold less than the mean). In this respect, this approach is similar to a recent and powerful proteomic approach, fluorescent two-dimensional differential gel electrophoresis (2-D DIGE), for which dedicated software such as the Extended Data Analysis module of DeCyder (GE Healthcare) is available.

An alternative method consists in standardizing data row by row using the classic zero-mean and unit-standard deviation technique.[22] This normalization has the advantage of being applicable either to the raw data matrix (VM-Center) or to the reliable one (VMR-Center). Moreover, this procedure is almost always implemented in the statistical package or the bioinformatic software available in proteomics.

Finally, the data matrices were submitted to a two-way normalization, the first being vertical (% vol), which makes gels comparable, the second horizontal, and this time making the proteins comparable.

Hierarchical Clustering Analysis (HCA): Methods and Tool. Once correctly preprocessed, the different data matrices may be submitted to two-way HCA; that is, HCA is applied independently to the columns and to the rows. The distance measurement selected for this evaluation is based either on the popular Pearson correlation (PE) coefficient ($1 - r$) or on the Euclidean metric (EU). The definition of these metrics is available in the Supporting Information. According to our own assessment, PE is quoted in about 33% of the proteomic literature (supplementary Table 1, Supporting Information). It is also advised in transcriptomics[14] when absolute spot hybridization levels are used. In fact, PE is less sensitive to the scale factor than EU so it may also be more appropriate to protein spot volumes. EU is used in about 54% of the proteomic literature (supplementary Table 1, Supporting Information). It is generally advised for normal data, and consequently, it seems more suited to logged ratio data[14] than raw volumes whose distribution is non-normal. EU is also more sensitive to missing values because this metric is differently weighted according to the number of MV contained in the profiles. Finally, both metrics (EU and PE) are affected more strongly by very abundant protein spot variation (minority) than low-abundance ones (majority). Consequently, profile normalization and MV imputation are essential. A third distance based on the popular Jaccard index (JA) is used exclusively for binary matrix computing.[23] The definition of this metric is available in the Supporting Information.

Concerning the aggregation procedures, the most commonly used ones: UPGMA (UP), complete linkage (CO), and Ward's methods (WA) have been tested in this study. A brief description of these terms is available in the Supporting Information. In transcriptomics, use of the complete linkage method gives generally better results[15] than UPGMA, but the latter is widely used in proteomics (52%) based on listed references (supplementary Table 1, Supporting Information). In contrast, Ward's method is almost never used with microarray data sets, whereas 20% of proteomic authors choose this aggregation procedure (supplementary Table 1, Supporting Information). The reasons for such differences between transcriptomics and proteomics are unknown and probably result simply from different practices without any precise knowledge of the advantages and limitations of each procedure. In proteomics, raw volumes are assumed to be not-normally distributed and "variance-vs-mean"-dependent. We can notably observe an increase of variance among the highest spot volumes. Data normalization is then generally beneficial[24] to any parametric approach such as the Ward's agglomeration method based on intragroup variance minimization or the UPGMA based on average calculation.

To evaluate each clustering methodology [a combination of data preprocessing steps 1 and 2 × distance × aggregation procedure], powerful and user-friendly software are needed. At present, transcriptomics offers a greater abundance of freely available data mining tools. The first and most popular one is certainly the Cluster+TreeView program by Eisen et al.[13] In addition to offering several clustering methodologies and the classic dendrogram display, this tool allows heat map visualization of the complete data matrix, a useful color representation in which each data point color is proportional to its value. Unfortunately, the clustering of binary data is not clearly designed, and Ward's method is not implemented in this tool, similarly to most other programs. Consequently, it is proposed here to use a new software program, PermutMatrix, which was also developed for microarray data but offers the entire panel of methods and visualization possibilities needed for this study. PermutMatrix version 1.8 is available free at http://www.lirmm.fr/~caraux/PermutMatrix/. Its convivial interface allows the five following steps for HCA:

(i) Uploading of a standard tab-delimited text file containing the data matrix

1. Columns represent samples (e.g., gels)

2. Rows represent protein spots

(ii) Setting of both clustering parameters, the distance and the aggregation procedure

1. Distance = Pearson (PE), Euclidean (EU), or Jaccard (JA)

2. Aggregation = UPGMA (UP), Complete (CO), or Ward (WA)

(iii) HCA application on the columns

(iv) HCA application on the rows

(v) Optimal reorganization of the trees using seriation methods (optional)

(vi) Clustering result visualization

1. Dendrogram of the samples

2. Dendrogram of the protein spots

3. Heat map of the clustered data matrix

Clustering Result Validation. In most proteomic studies, HCA is first used to construct the dendrogram of the samples. Then the biochemist generally accepts at least one of the grouping results and interprets it based on his prior biological knowledge. Unfortunately, the choice of the optimal clustering solution is sometimes subjective and, consequently, debatable and sometimes inadequate. To validate a clustering result, its quality may also be evaluated statistically using various "internal" criteria, but these are known to be biased because they are highly correlated with the clustering algorithm.[14] Thus, "external" criteria are generally preferred when methodologies have to be compared.[14,25] Therefore, it is proposed to use an external quality criterion based on the most reasonable hypothesis that can be intuitively raised here: since experimental variability is generally expected to be lower than biological variability, natural partition of the data results in grouping the experimental replications in classes of n_{epx} gels. Then, a useful quality metric, the *F*-measure,[25] may be computed to compare each entire clustering solution to the natural partition. The *F*-measure is a combination of the "precision" and the "recall" assessment, where precision expresses the proportion of well-classed objects (e.g., gels) for each cluster and recall expresses the proportion of well-classed objects for each class. The *F*-measure is computed as described in the Supporting Information,[25] with this positive number taking 1 as maximum value when the clustering solution is equal to the natural partition. Finally *F*-measure is actually relevant because all replicate gels come from different batches. This condition is absolutely essential. Additionally, all the dendrograms constructed under PermutMatrix were positively validated by STATISTICA version 6.1 (StatSoft, France) which is a widely recognized and easy-to-use statistical analysis software. The topology of these dendrograms may also be a good indicator of the clustering quality. In fact, a well-balanced tree generally reflects biological reality better than an unbalanced one, plots in stairs being a good example.

Clustering Method Evaluation. First, to evaluate the global effect of the clustering methodology on the clustering results, the *F*-measure was calculated for each clustering solution. All results were presented as means between both data sets.

Second, the effect of the missing value imputation method was compared using suitable clustering methodologies, that is, those giving the highest *F*-measure.

Finally, the global capability of the clustering methodology was evaluated from a practical point of view, including visual diagnostics, cluster extraction possibility, and conviviality.

Results and Discussion

Effect of the Algorithm and the Data Normalization. We evaluated the effect of the global methodology on the clustering result of the samples. The selected algorithms resulted from a combination of a distance metric (EU or PE) and an aggregation procedure (UP, CO, or WA). They were executed on the reliable matrix of each data set (DS1 and DS2) after one of the proposed volume normalization procedures: (i) VMR (raw volume without normalization), (ii) VMR-Ratio, and (iii) VMR-Center. Bar graphs in Figure 1 depict the *F*-measure levels obtained for all possible clustering solutions.

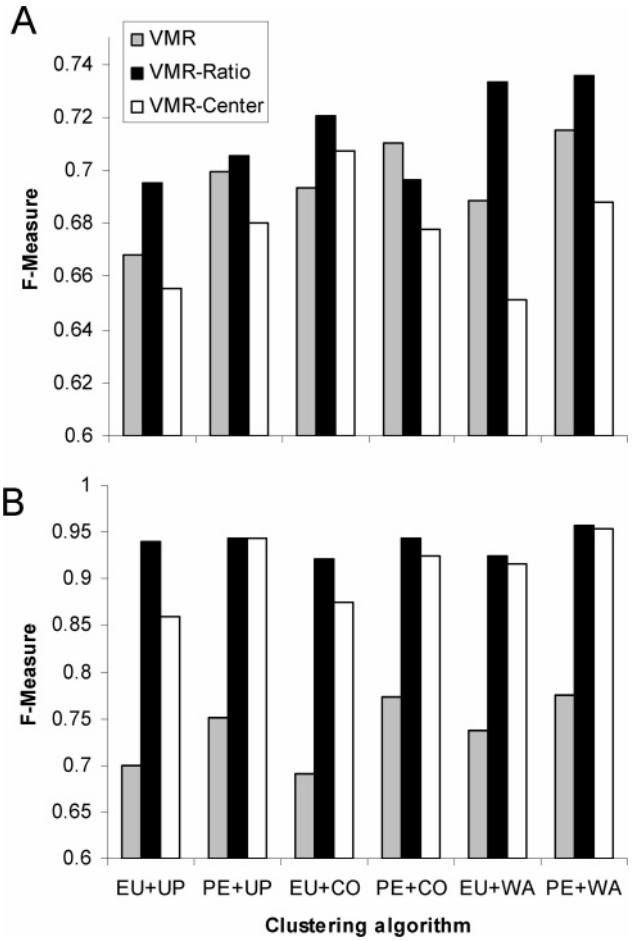

Figure 1. Effects of the algorithm and the data normalization methods on the clustering results validated by *F*-measure. Bar graphs depict the *F*-measure levels obtained for each clustering solution. (A) Hierarchical clustering analysis (HCA) applied to the reliable matrix (VMR) of the first data set (DS1) without data normalization (gray bars), after logged ratio transformation (black bars), or 0-mean and 1-standard deviation normalization (white bars). The clustering algorithms result from a combination of a distance metric (EU = Euclidean or PE = Pearson correlation) and an aggregation procedure (UP = UPGMA, CO = Complete, or WA = Ward). (B) Similar analysis made with the second data set (DS2).

1. Data Normalization Effect. The effect of the data normalization method was important, but a significant interaction was observed between data normalization and data set (Table 1A). The Ratio-based approach (VMR-Ratio) showed a higher performance for both data sets than the center-based (VMR-Center) and than the no-normalization methods (VMR), and this effect was higher for DS2 than for DS1. The worst method was VMR-Center for DS1 and VMR for DS2.

2. Clustering Algorithm Effect. The effect of the clustering algorithm was also important, but a significant interaction was observed between clustering algorithm and data set (Table 1B).

Table 1. Effects of the Data Normalization Method (A) and the Clustering Algorithm (B) on the Clustering Results for Both Data Sets (DS1 and DS2) Validated by the *F*-Measure[a]

A

data set (DS)	data normalization method (DNM)		
	VMR	VMR-Ratio	VMR-Center
DS1	0.696	**0.714**	0.677
DS2	0.738	**0.938**	0.912

B

data set (DS)	clustering algorithm (CA)					
	EU+UP	PE+UP	EU+CO	PE+CO	EU+WA	PE+WA
DS1	0.673	0.695	0.707	0.695	0.691	**0.713**
DS2	0.833	0.879	0.829	0.880	0.859	**0.896**

[a] (A) Three data normalization methods (DNM) are evaluated: (i) VMR (raw volume without normalization), (ii) VMR-Ratio (logged ratio transformation), and (iii) VMR-Center (0-mean and 1-standard deviation normalization). (B) The six clustering algorithms (CA) tested are a combination of one distance [Pearson (PE) or Euclidean (EU)] plus one aggregation procedure [UPGMA (UP), Complete (CO), or Ward (WA)]. For each data set, highest *F*-measures (corresponding to best clustering strategies) are indicated in bold.

Nevertheless, the PE+WA algorithm was the best solution whatever the data set used, even if the difference between this algorithm and the others was sometimes low. Consequently, this result needs to be confirmed in further studies. Taken individually, PE was a better distance metric than EU in most of the cases, but this difference was also weak. Similarly, Ward's aggregation procedure appeared noticeably more powerful than the others.

3. Validation of the *F*-Measure by Biological Considerations. Additionally, it was interesting to note the high correlation between this *F*-measure, reflecting the quality of the replication grouping, and a biologically relevant grouping of the samples. Thus, in DS1 analysis, only the clustering methodologies giving an *F*-measure above about 0.7 led to clear separation of strains 1−6 from strains 7−12 in a balanced tree (Figure 2). This result was consistent with the phylogenetic analyses based on genomic data which showed that the 13 recognized serovars identified in the *L. monocytogenes* species were distributed between two different lineages, with serovars 1/2b and 4b strains in particular found in lineage I and serovar 1/2a strains found in lineage II.[26,27] Identification and validation of the protein markers splitting the two lineages is under way. Concerning DS2, a similar phenomenon appeared for *F*-measure above about 0.9, leading to a clear distinction of 30-month-old rats from others and a weaker distinction between 7- and 18-month-old rats (Figure 3). ANOVA analysis previously performed on this data set showed a similar result.[17] Indeed, mean differences observed for the significant differentially expressed proteins between 7- vs 18-, 18- vs 30-, and 7- vs 30-month-old rats were respectively 77%, 92%, and 107%, whereas the mean coefficients of experimental and biological variations were 25.3% (n_{exp} = 3) and 29.3% (n_{bio} = 5), respectively. Consequently, the biological differences detected by both ANOVA and clustering were higher than technical or biological variabilities, which validates our analysis.

In conclusion, even if our results were shown with only two data sets, we have demonstrated here the weakness of popular methodologies, (e.g., EU+UP) which surprisingly gave the lowest *F*-measure (Figure 1), the worst clustering results, and the most unbalanced trees (data not shown). Data normalization was also needed, although it is almost never mentioned in the publications. Thus, the application of the PE+WA algorithm to the VMR-Ratio gave the best *F*-measure in almost all the situations, leading to well-balanced and easily legible dendrograms (Figures 2 and 3). Unfortunately, this data normalization was only applicable on a reliable (and consequently generally reduced) data matrix containing no missing value.

Effect of the Missing Value Imputation. In this second comparative study, the overall missing value imputation techniques were evaluated based on the algorithm that performed best, adapted to each data set: JA+WA was selected for HCA of the binary matrix (VM-Bin) and PE+WA was used for the other matrices. Table 2 summarizes these methods and the maximum *F*-measure levels obtained for each best clustering solution. Interestingly, HCA applied to the binary matrix of the first data set led to an almost perfect grouping of replicate gels (*F*-measure = 0.97). Here, the effect of absent/present biomarkers seemed predominant compared to the quantitative values. Nevertheless, clustering of the VM-Zero-Center was equally efficient (*F*-measure = 0.91), with this matrix integrating the benefits of both quantitative and binary information. Because of 36% missing values, the clustering software was unable to process the raw matrix VM-Center. Finally, HCA of the VMR-Ratio gave the worst result, notably because only 140 out of a total of 599 proteins were taken into account. Thus, we may suppose these 140 proteins are not only common to all strains but most of them are housekeeping proteins characterized by similar expression profiles. Concerning the second data set, HCA applied to the binary matrix was this time clearly uninformative; 4% missing values seemed inadequate to contain pertinent or even discriminating information. The process applied to the three other normalized matrices (VM-Center, VM-Zero-Center, and VMR-Ratio) gave quite similar and satisfying results (0.93, 0.94, and 0.95, respectively), with the maximum *F*-measure value being obtained with PE+WA algorithm.

In conclusion, each data set is specific and, consequently, may need one or several suitable treatments. Here, both data sets were of interest and complementary for our study. DS1 was mostly characterized by genetic differences between samples involving notably numerous absent/present markers. A binary approach was informative and powerful in this case. On the contrary, DS2 was mainly constructed for differential expression analysis where small quantitative differences were expected. Here, a ratio-based approach seemed more sensitive for revealing proteins of interest.

Major HCA Assets. Once the samples are correctly grouped, either automatically using the well-adapted clustering methodology or manually leaving samples naturally arranged in order (e.g., in a kinetic study), it may be interesting to group proteins showing similar expression profiles. To this end, two-way HCA is a suitable preliminary data mining approach. Heat map visualization of the binary (Figure 2) or the quantitative (Figure 3) clustered data matrices gives powerful diagnostic information. Groups of coregulated proteins can be rapidly identified, either by their absence/presence pattern as shown in clusters C1 and C2 of Figure 2, or their quantitative difference, in clusters C1 and C2 of Figure 3. In this latter figure, only the data normalization techniques recommended previously (-Ratio or -Center) allowed this powerful visualization and the possibility for extraction of complex patterns. In a similar manner, it may be interesting to identify natural or

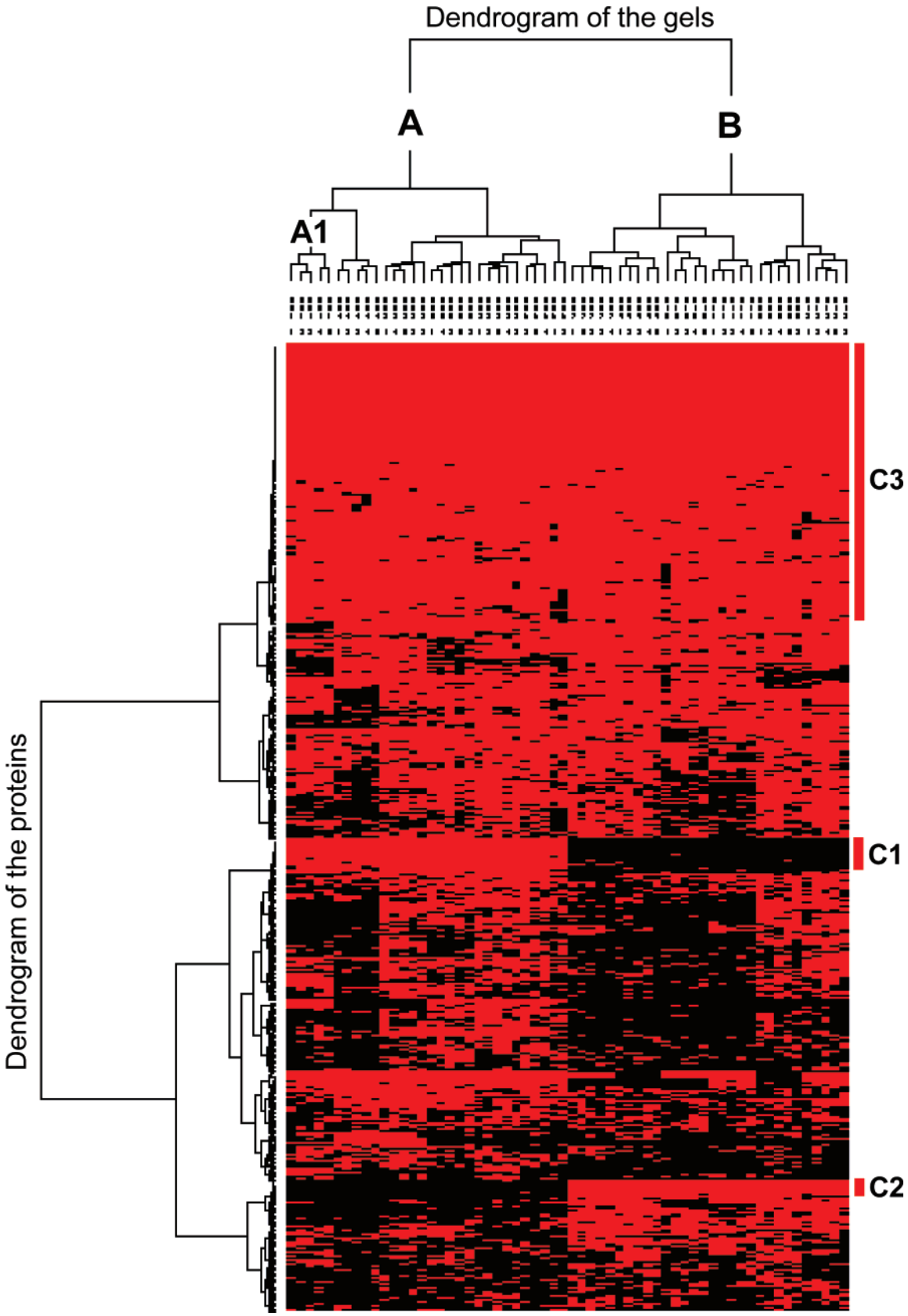

Figure 2. Two-way hierarchical clustering analysis (processed with PermutMatrix according to the Jaccard index and Ward's aggregation method) of the first data set (DS1, 599 proteins × 60 gels) after a binary transformation (VM-Bin). Heat map representation of the clustered data matrix in which each red cell represents an existing protein value (1) and each black cell represents a missing protein value (0). The dendrogram of the gels shows a balanced tree in which strains 1−6 are clearly separated (parent cluster [A]) from strains 7−12 (parent cluster [B]). Moreover, replicated gels are mostly grouped in separate subclusters (child cluster [A1], for example) leading to the higher *F*-measure obtained for this data set (0.97). The dendrogram of the proteins allows visual selection of coregulated proteins, for example, biomarkers present only in a predefined selection of strains: cluster [C1] of 22 proteins present only in the strains 1−6 and cluster [C2] of 12 proteins present only in the strains 7−12. In contrast, cluster [C3] is composed of 174 "reliable" proteins with relatively few missing values; thus, this cluster needs to be reanalyzed in a quantitative manner.

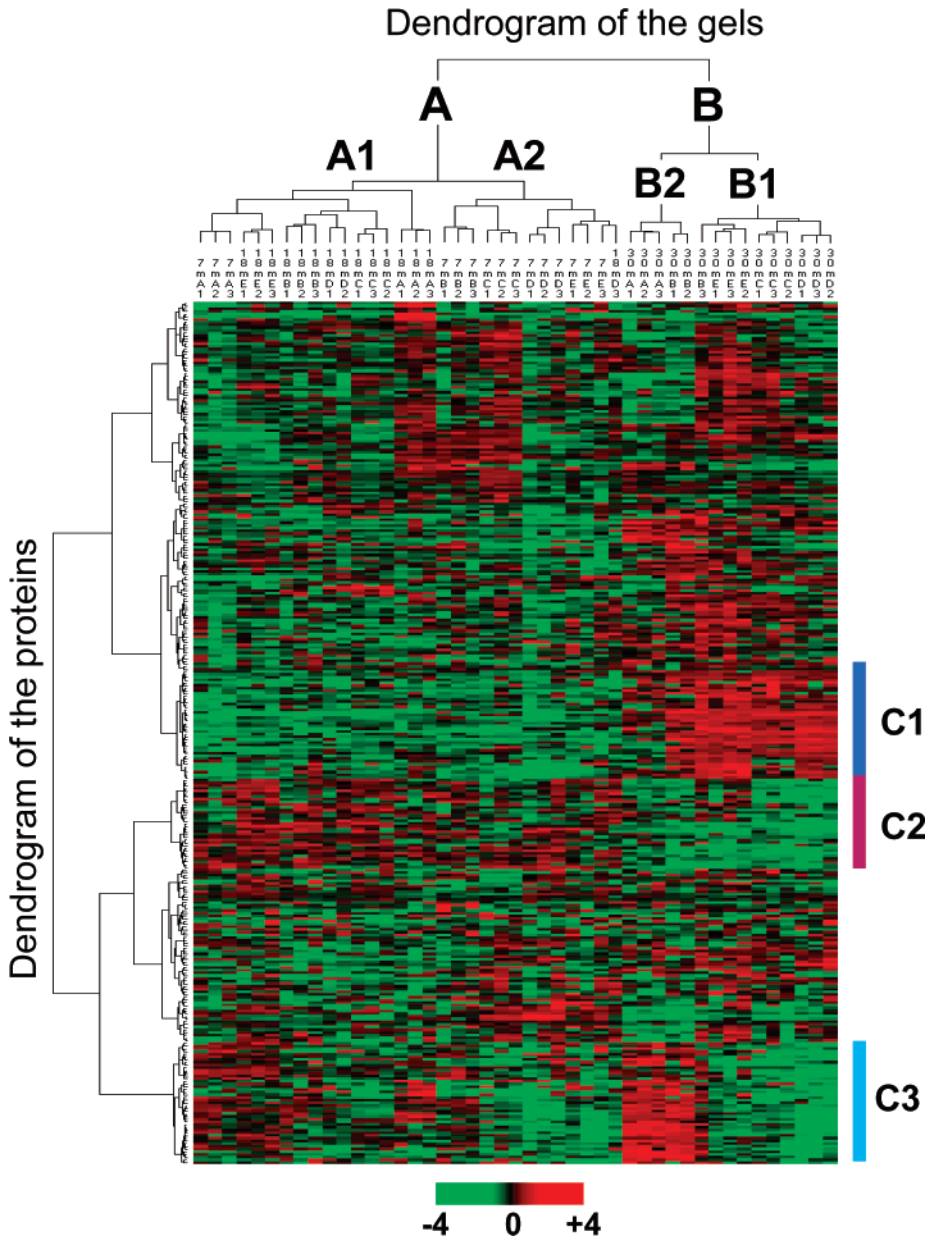

Figure 3. Two-way hierarchical clustering analysis (processed with PermutMatrix according to the Pearson distance and Ward's aggregation method) of the second data set (DS2, 318 proteins × 45 gels) after a logged-ratio transformation (VMR-Ratio). Heat map representation of the clustered data matrix in which each colored cell represents a protein value according to the color scale at the bottom of the figure. The dendrogram of the gels depicts a balanced tree in which all 30-month-old rats are clearly separated [B] from 7- and 18-month-old rats grouped in a common cluster [A] with a weak distinction between 7- (cluster [A2]) and 18-month-old rats (cluster [A1]). Clusters of coregulated proteins [C1] or [C2] illustrate this phenomenon, respectively, in terms of up-expression or down-expression in 30-month-old rats. Among the five rats aged 30 months, two (cluster [B2], characterized by [C3]) appear to act like outliers. Additionally, replicated gels are mostly grouped in separate subclusters leading to the higher *F*-measure obtained for this data set (0.95).

artificial (mainly due to experimental errors) groups of outliers, among the samples or the proteins. Thus, HCA of DS2 highlighted clearly two rats inside the "30-months" group (Figure 3, clusters B2/C3), whereas, on the contrary, few differences were observed inside the 7- and 18-months one. This first understanding of the data set, without any *a priori* knowledge, may orient future data analysis strategy such as new model construction for ANOVA or data filtering before further processing.

General Conclusions. As already shown in a previous work,[3] there is much similarity between transcriptomic and proteomic data analysis strategy, and it needs to be exploited. Consequently, it is clear that bioinformatic tools, mainly developed for microarrays, could be adapted to 2-D PAGE-based studies. Unfortunately, this is not yet generally true and, in any case, needs careful attention. The most popular methodologies are sometimes weak. Here, we recommend PermutMatrix, a new free software implementing HCA methodology. We have shown that it is well-adapted to the present proteomic data sets. In addition, we have highlighted specific features of proteomic data: unlike in transcriptomics, where Ward's aggregation method is almost never employed, this procedure gives the best

Table 2. Effect of the Missing Value (MV) Imputation Method on the Clustering Results Validated by the *F*-Measure[a]

data set	matrix	size	MV (%)	best algorithm		*F*-measure MAX
				distance	aggregation	
DS1	VM-Center	599 × 60	36	-	-	-
	VM-Bin	599 × 60	0	Jaccard	Ward	**0.97**
	VM-Zero-Center	599 × 60	0	Pearson	Complete	0.91
	VMR-Ratio	140 × 60	0	Pearson	Ward	0.73
DS2	VM-Center	341 × 45	4	Pearson	Ward	0.93
	VM-Bin	341 × 45	0	Jaccard	Ward	0.39
	VM-Zero-Center	341 ×45	0	Pearson	Ward	0.94
	VMR-Ratio	318 ×45	0	Pearson	Ward	**0.95**

[a] Four methods are proposed: (i) no imputation (VM-Center), (ii) binary transformation (VM-Bin), (iii) replacement by zero (VM-Zero-Center), and (iv) replacement by mean (if available) of replicate values (VMR-Ratio). For each imputation technique, the corresponding matrix, reduced or not, is submitted to the HCA method (best algorithm and normalization procedure) giving the maximum *F*-measure.

clustering results in both these proteomic studies, notably, when associated with the Pearson-based distance metric. Similarly, the clustering of binary data appears particularly informative either to identify visually a protein spot strictly absent in one condition or to highlight gels with too many MVs. It is also important to note that the presence of replicate gels is an absolute prerequisite to select objectively the best clustering solutions among several, all of which give results that are sometimes quite different. This is especially valuable if technical replicate gels result from different batches, from extraction to staining. Interestingly, this well-tuned HCA approach gives the most biologically relevant grouping of the samples. Biological replicates are, thus, essential to strengthen the concluding remarks about the population of concern.[28] In conclusion, the two-way clustering approach is easy to implement, but special care must be taken with respect to data normalization, missing value imputation (which depends on data set specific features), and cluster validation (by "external" criteria). HCA may also be advantageously combined with other multivariate exploratory techniques such as principal components analysis (PCA) or *K*-means and Self-Organizing Maps (SOM)[29] clustering methods.

Acknowledgment. We thank Drs. I. Cassar-Malek, J. F. Martin, C. Jurie, and C. Bernard for constructive and helpful discussions. This work was supported by the AGENAE program. AGENAE is a national French program related to structural and functional genomics applied to animal science. The authors also thank the ≪ Commissariat à l'aménagement et au Développement Economique du Massif Central ≫ (France) for its financial support.

Supporting Information Available: Supplementary Table 1 depicts the 28 recent proteomic publications analyzed in this study. In addition, a brief description of clustering terms is proposed. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Biron, D. G.; Brun, C.; Lefevre, T.; Lebarbenchon, C.; Loxdale, H. D.; Chevenet, F.; Brizard, J. P.; Thomas, F. The pitfalls of proteomics experiments without the correct use of bioinformatics tools. *Proteomics* **2006**, *6*, 5577−5596.

(2) Chang, J.; Van Remmen, H.; Ward, W. F.; Regnier, F. E.; Richardson, A.; Cornell, J. Processing of data generated by 2-dimensional gel electrophoresis for statistical analysis: missing data, normalization, and statistics. *J. Proteome Res.* **2004**, *3*, 1210−1218.

(3) Meunier, B.; Bouley, J.; Piec, I.; Bernard, C.; Picard, B.; Hocquette, J. F. Data analysis methods for detection of differential protein expression in two-dimensional gel electrophoresis. *Anal. Biochem.* **2005**, *340*, 226−230.

(4) Iwadate, Y.; Sakaida, T.; Hiwasa, T.; Nagai, Y.; Ishikura, H.; Takiguchi, M.; Yamaura, A. Molecular classification and survival prediction in human gliomas based on proteome analysis. *Cancer Res.* **2004**, *64*, 2496−2501.

(5) Harris, R. A.; Yang, A.; Stein, R. C.; Lucy, K.; Brusten, L.; Herath, A.; Parekh, R.; Waterfield, M. D.; O'Hare, M. J.; Neville, M. A.; Page, M. J.; Zvelebil, M. J. Cluster analysis of an extensive human breast cancer cell line protein expression map database. *Proteomics* **2002**, *2*, 212−223.

(6) Chevalier, F.; Martin, O.; Rofidal, V.; Devauchelle, A. D.; Barteau, S.; Sommerer, N.; Rossignol, M. Proteomic investigation of natural variation between Arabidopsis ecotypes. *Proteomics* **2004**, *4*, 1372−1381.

(7) Appel, R.; Hochstrasser, D.; Roch, C.; Funk, M.; Muller, A. F.; Pellegrini, C. Automatic classification of two-dimensional gel electrophoresis pictures by heuristic clustering analysis: a step toward machine learning. *Electrophoresis* **1988**, *9*, 136−142.

(8) Vohradsky, J. Adaptive classification of two-dimensional gel electrophoretic spot patterns by neural networks and cluster analysis. *Electrophoresis* **1997**, *18*, 2749−2754.

(9) Jimenez, C. R.; Stam, F. J.; Li, K. W.; Gouwenberg, Y.; Hornshaw, M. P.; De, Winter, F.; Verhaagen, J.; Smit, A. B. Proteomics of the injured rat sciatic nerve reveals protein expression dynamics during regeneration. *Mol. Cell. Proteomics* **2005**, *4*, 120−132.

(10) Culp, W. D.; Neal, R.; Massey, R.; Egevad, L.; Pisa, P.; Garland, D. Proteomic analysis of tumor establishment and growth in the B16−F10. mouse melanoma model. *J. Proteome Res.* **2006**, *5*, 1332−1343.

(11) Dowsey, A. W.; Dunn, M. J.; Yang, G. Z. The role of bioinformatics in two-dimensional gel electrophoresis. *Proteomics* **2003**, *3*, 1567−1596.

(12) Bensmail, H.; Haoudi, A. Postgenomics: proteomics and bioinformatics in Cancer Research. *J. Biomed. Biotechnol.* **2003**, *2003*, 217−230.

(13) Eisen, M. B.; Spellman, P. T.; Brown, P. O.; Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 14863−14868.

(14) D'haeseleer, P. How does gene expression clustering work? *Nat. Biotechnol.* **2005**, *23*, 1499−1501.

(15) Gibbons, F. D.; Roth, F. P. Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res.* **2002**, *12*, 1574−1581.

(16) Caraux, G.; Pinloche, S. PermutMatrix: a graphical environment to arrange gene expression profiles in optimal linear order. *Bioinformatics* **2005**, *21*, 1280−1281.

(17) Piec, I.; Listrat, A.; Alliot, J.; Chambon, C.; Taylor, R. G.; Bechet, D. Differential proteome analysis of aging in rat skeletal muscle. *FASEB J.* **2005**, *19*, 1143−1155.

(18) Cox, B.; Kislinger, T.; Emili, A. Integrating gene and protein expression data: pattern analysis and profile mining. *Methods* **2005**, *35*, 303−314.

(19) de Brevern, A. G.; Hazout, S.; Malpertuy, A. Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. *BMC Bioinf.* **2004**, *5*, 114.

(20) Vohradsky, J.; Janda, I.; Grunenfelder, B.; Berndt, P.; Roder, D.; Langen, H.; Weiser, J.; Jenal, U. Proteome of Caulobacter crescentus cell cycle publicly accessible on SWICZ server. *Proteomics* **2003**, *3*, 1874−1882.

(21) Gion, J. M.; Lalanne, C.; Le, Provost, G.; Ferry-Dumazet, H.; Paiva, J.; Chaumeil, P.; Frigerio, J. M.; Brach, J.; Barre, A.; de Daruvar, A.; Claverol, S.; Bonneu, M.; Sommerer, N.; Negroni, L.; Plomion, C. The proteome of maritime pine wood forming tissue. *Proteomics* **2005**, *5*, 3731−3751.

(22) Satoh, M.; Haruta-Satoh, E.; Omori, A.; Oh-Ishi, M.; Kodera, Y.; Furudate, S.; Maeda, T. Effect of thyroxine on abnormal pancreatic proteomes of the hypothyroid rdw rat. *Proteomics* **2005**, *5*, 1113−1124.

(23) Zivy, M.; el Madidi, S.; Thiellement, H. Distance indices in a comparison between the A, D, I and R genomes of the Triticeae tribe. *Electrophoresis* **1995**, *16*, 1295−1300.

(24) Gustafsson, J. S.; Ceasar, R.; Glasbey, C. A.; Blomberg, A.; Rudemo, M. Statistical exploration of variation in quantitative two-dimensional gel electrophoresis data. *Proteomics* **2004**, *4*, 3791−3799.

(25) Handl, J.; Knowles, J.; Kell, D. B. Computational cluster validation in post-genomic data analysis. *Bioinformatics* **2005**, *21*, 3201−3212.
(26) Zhang, C.; Zhang, M.; Ju, J.; Nietfeldt, J.; Wise, J.; Terry, P. M.; Olson, M.; Kachman, S. D.; Wiedmann, M.; Samadpour, M.; Benson, A. K. Genome diversification in phylogenetic lineages I and II of Listeria monocytogenes: identification of segments unique to lineage II populations. *J. Bacteriol.* **2003**, *185*, 5573−5584.
(27) Doumith, M.; Cazalet, C.; Simoes, N.; Frangeul, L.; Jacquet, C.; Kunst, F.; Martin, P.; Cossart, P.; Glaser, P.; Buchrieser, C. New aspects regarding evolution and virulence of Listeria monocytogenes revealed by comparative genomics and DNA arrays. *Infect. Immun.* **2004**, *72*, 1072−1083.
(28) Karp, N. A.; Spencer, M.; Lindsay, H.; O'Dell, K.; Lilley, K. S. Impact of replicate types on proteomic expression analysis. *J. Proteome Res.* **2005**, *4*, 1867−1871.
(29) Herrero, J.; Dopazo, J. Combining hierarchical clustering and self-organizing maps for exploratory analysis of gene expression patterns. *J. Proteome Res.* **2002**, *1*, 467−470.

PR060343H