MMS: A "R" Package for Metabolomic Markers Search in Stress Response Studies

G. Cardinali*, P. Rellini, C. Pelliccia, L. Antonielli and F. Fatichenti

DBA - Microbiology, University of Perugia. Via Borgo XX Giugno 74, I-06121 Perugia, Italy

Abstract: In stress response studies, metabolomics tries to individuate spectral areas or metabolites whose response to a specific stress is particularly significant. This evaluation is normally carried out with a series of tedious and time consuming comparisons of spectral areas obtained from cells subject to different stress intensities. MMS is a package written in "R" language able to individuate which part of metabolomic spectra have a statistically significant response giving as output the diagrams of the R^2 and of the slope calculated from regression analysis between a vector describing the intensity of the stress and a matrix containing all spectra from differently challenged cells.

INTRODUCTION

Metabolomics is the science that studies the whole metabolic composition of cells with a series of techniques like the Fourier Transform InfraRed Spectroscopy (FTIR), the Capillary Electrophoresis (CE) and several others. The output of most metabolomic techniques is some sort of spectrum with various types of descriptors on the abscissa, such as wavenumbers, wavelengths, retention time etc. Intensity or absorbance are normally reported on the ordinate.

The metabolomic analysis has been proposed and is currently used in a large spectrum of fields, among which taxonomy [1-6], strain characterization [7-10], isolate dereplication [11] and the study of the effects caused by different biological, physical or chemical agents, which will be hereinafter referred to as stress response [12-13]. Stress response is one of the most promising areas of modern biology for its theoretical and applicative perspectives. In fact, from the point of view of basic research, although we can depict completely the genomes and also the proteomes, a dynamic observation of how the different cells work in different conditions is still poorly understood. By a practical viewpoint, stress response studies are giving the possibility of early and effective monitoring of risky or dangerous situations spanning from the clinical applications [14-18] to the environmental protection [19].

In almost all cases the study of stress response focuses on the search for metabolomic markers, i.e. spectral areas whose response to the challenge is particularly intense or significant. Since stress response studies are particularly effective when based on complex experimental designs, this investigations tends to become particularly difficult and time consuming if no informatic support is provided.

This article describes a package written in the free statistical environment "R" (http://cran.r-project.org/) named MMS as acronym of Metabolomic Markers Search. This software compares all spectral areas with a reference vector reporting the intensities of the stressing conditions, such as the exposition time, concentration or intensity of the stress under study. For each single area, a linear or a logarithmic regression is calculated giving as output the R^2 and the curve slope. The former parameter indicates how trustful is the regression, whereas the second evaluates the strength of the response to the challenge. These graphs represent "response diagrams" to immediately visualize the most interesting spectral regions on which further analyses should be carried out. Other auxiliary functions are available to facilitate additional investigations aiming to individuate significant and robust metabolomic markers.

METHODS

Strains and Growth Conditions

The yeast employed in this study is the type strain of the species *Saccharomyces cerevisiae*: DBVPG 6173, obtained from the Industrial Yeasts Collection DBVPG (http://www.agr.unipg.it/dbvpg/home.html), which was grown on YEGA (1% Yeast Extract, 1% Glucose, 1.7% Agar) at 25°C for 24h. Medium components were obtained from Difco (USA).

Stress Conditions

Cells were harvested with a platinum 1.5 mm diameter loop. One loopful was suspended in 105 µl distilled sterile water to obtain an optimal concentration for the FT-IR analyses. The suspension was added with sulfur dioxide to reach the two experimental conditions: 200 mg/l and 400 mg/l. The time course test was carried out by exposing cell suspensions to the appropriate SO_2 concentration for six time periods: 0h, 1h, 2h, 4h, 6h and 24 h, the first being considered the resting condition, i.e. the state in which no stress was exerted on cells.

FTIR Analysis

FT-IR measurements were performed in transmission mode. All spectra were recorded between 4000 and 400 cm^{-1} with a TENSOR 27 FT-IR spectrometer (BRUKER Optics GmbH, Ettlingen, Germany) with 4 cm^{-1} spectral resolution. Each spectrum resulted from 64 scan samplings. The software OPUS version 6.5 (BRUKER) was used for the primary analysis of the spectra (first derivative and baseline correction). Further statistical analyses were carried out with the free source "R" package from CRAN (http://cran.r-project.org/src/contrib/PACKAGES.html).

*Address correspondence to this author at the DBA - Microbiology, University of Perugia. Via Borgo XX Giugno 74, I-06121 Perugia, Italy;
E-mail: gianlu@unipg.it

1874-1363/08

Experimental Design

MMS has been designed considering a common experimental design in which the same organism is subject to different challenging situations such as the increasing concentration or different times of exposition to a stressing agent. This design will produce as many spectra as the conditions investigated plus the control condition with no challenge. The number of descriptor depends on the technique employed and on the experimental setting.

In the case taken as example, the organism is the yeast *Saccharomyces cerevisiae*, the stressing agent is the sulfur dioxide and the conditions investigated are the following times of exposition: 1h, 2h, 4h, 6h and 24h. An additional spectrum acting as reference and control was obtained from cells without any addiction and therefore considered as time 0h.

The descriptors are the wavenumbers spanning from 4000 cm^{-1} to 600 cm^{-1}. The spectrophotometer recorded the vibrational intensity every 1.926 cm^{-1}, producing 1194 values, hereinafter referred to as descriptors.

The output matrix consisted of seven objects (the seven times including the control) and 1194 descriptors.

The reference vector is a simple series of data describing the conditions imposed to the experiment. In the case presented it is a string of seven values reporting the hours of exposition to the SO_2: 0, 1, 2, 4, 6, 24.

RESULTS

Presentation of the Algorithms Included in MMS

MMS includes nine functions subdivided as importing functions, MMS itself and the auxiliary functions to manage data after the primary analysis. These functions are described below orderly as they would be used in a normal analysis to search metabolomic markers (Figs. **1**,**2**).

The function ***ref*** is used to import the reference vector, which can be generated with any test processing program or with a calculation sheet, provided that no header is included and that the file be saved as tab spaced text (extension .txt).

The matrix containing the spectra, imported with the function ***spec***, should include the header, reporting the descriptors' labels, and a first column with the objects' labels.

The core of the package is ***mms***, a function searching spectral regions in which the change of intensity is somehow related to the different challenging conditions (in our case the exposition time), by regression analysis. In order to carry out a regression study, two vectors of the same size are required. In our case one vector (say *x*) is *ref*, whereas the other (*y*) can be the data series of each descriptor (in our case the seven intensities of each wavenumber). Since data from a spectral area as narrow as 2 cm^{-1} can be difficult to interpret and scarcely significant, the algorithm can calculate the *y* vector as average of several vectors of contiguous descriptors. The analyst can decide how many descriptors should be considered together by setting the *sf* parameter, whose default, and minimum possible value, is 2. After this preliminary step, the algorithm works iteratively in a sliding window manner as described below.

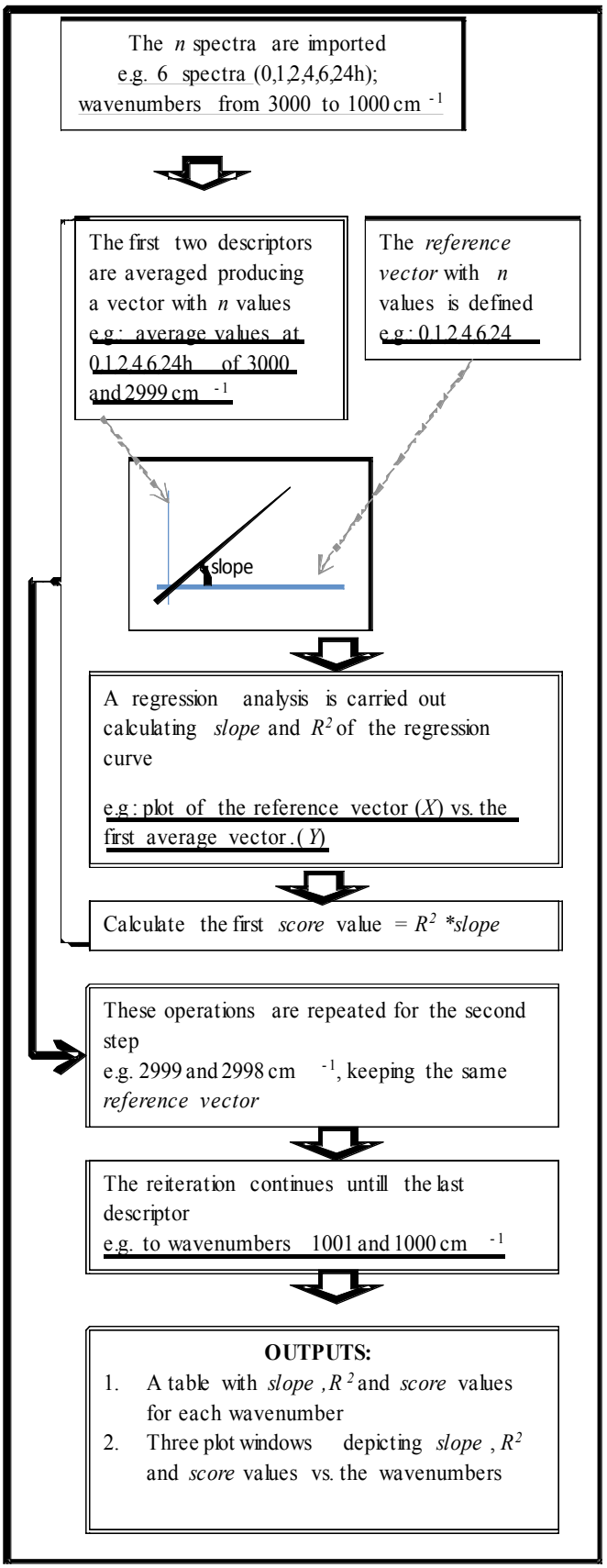

Fig. (1). Flow chart of the MMS algorithm. The MMS algorithm is illustrated in general and in the specific case of FTIR analysis, in which the wavenumbers are the descriptors. The same algorithm can be applied to any other metabolomic technique.

1. The average vector (*y*) is calculated considering the descriptors spanning from 1 to *sf*.
2. A regression analysis is carried out between *ref* and the average vector, calculating both the slope of the regression curve and the R^2.

In the second reiteration, the (*y*) is calculated with the descriptors spanning from 2 to *sf*+1. In general, the window considered in each i^{th} step is from *i* to *sf*+*i*-1

The slope values are normalized in order to range from -1 to +1 with the formula 1.

Formula 1 $Sn= (S-Smin) / (Smax-Smin)$

where *Sn* is the normalized value, *S*, *Smin* and *Smax* are respectively the slope before the normalization, the minimum and the maximum slope values in the matrix. Since, by definition, the R^2 values range from 0 to 1it is now possible to calculate a *score* for each descriptor (or groups of descriptors) as product of the R^2 and of the slope.

The result matrix contains for each descriptor all three values of R^2, slope and score which indicate respectively the quality of the regression, the strength of the response, and a synthesis of the two former indicators.

Finally, each series of data is used to produce a plot in a separate window.

The algorithm includes the possibility to consider a logarithmic regression; in this case the *ref* vector is transformed according to formula 2:

Formula 2 $vl = ln(v+0.0001)$

In which *v* and *vl* represent respectively the original value of the *ref* vector and its natural logarithm. The 0.0001 value is added to avoid that the argument of the logarithm be null.

Auxiliary functions. After *mms* has been carried out, it is possible to export the result matrix with the *export* function. The *odi* (optimal descriptor identification) function allows to extract from the original matrix all those descriptors whose score is higher, as absolute value, than a threshold defined by the analyst. The threshold default is 0.75. In case the threshold is too high, a warning message appears as output.

In order to compare different experiments, the score obtained with the *mms* function can be stored in memory with the *loadN* function. Scores can be stored in each of the four memories available, simply writing the number 1 (or 2 or 3 or 4) after "load", e.g. *load2* for memory 2 etc.

Scores stored in memory can be compared with the *compare* function in which the analyst should state which scores should be processed, so *compare(r1,r3)* means that the score loaded in memory 1 will be compared with those in memory3. The algorithm simply subtracts the values of the second argument from the first, restituting a "comparison plot" with subtraction values in the ordinate axis and descriptors in the abscissa.

The *dif* function is the only not requiring the reference vector, it works as follows: the value of the control are subtracted from those of all other conditions. Then data of each single descriptor are averaged. The function yields a plot with these average differences from the control condition reported in the ordinate, as usual the abscissa reports the descriptors.

Finally, a very simple *help* function gives a rapid overview of the available functions and the basic syntaxes.

Validation of the Package

The MMS package was validated with data obtained with an experiment in which yeast cells were subject to the action of 200 mg/l sulfur dioxide for six different time periods spanning from 1 to 24 hours (0, 1, 2, 4, 6 and 24h). In order to test the *compare* function, a similar experiment with 400 mg/l SO_2 was carried out. In many stress response experiments analyzed with metabolomic tools only the control and one stressing conditions are considered. In our case, a quantitative approach was undertaken, with six time points, considered as a reasonable minimum to calculate a regression. However, since the points were not particularly numerous, we carried out the experiments with both Spearman and Pearson correlation, obtaining little if any difference. Moreover, introducing all repetition data (18 spectra) did not produce a remarkable difference from the use of the averages, as illustrated below.

The first elaboration to validate the software considered all the repetitions as separate objects, obtaining a matrix with 18 objects and all wavenumbers from 3000 cm^{-1} to 700 cm^{-1} with a total of 1194 descriptors.

Applying the *mms* function in the linear regression mode the three windows reported as Fig. (**2a-c**) were obtained. A further step was to set the regression in the logarithmic mode obtaining the other three panels of Fig. (**2**). The two analyses are in agreement regarding the regions from 1700 to 700 cm^{-1} characterized by high levels of R^2 (Fig. **2a**,**d**), agreement is also visible in the two major peaks of the slope graph (Fig. **2b**,**e**) in the regions around 1660 cm^{-1} and 1702 cm^{-1}. Interestingly, these two peaks, although clearly prominent differ in the slope value, which is close to 1 using the linear regression, whereas is around 0.6 when the logarithmic regression is used. Similarly, the graph area with negative values shows peaks at the corresponding wavenumbers, although with different levels of slope. The "score graphs" (Fig. **2c**,**f**) are a synthesis of the corresponding slope and R^2 graphs, with a major change in the fact that only a prominent positive peak is present in the logarithmic score plot (1660 cm^{-1}), whereas the linear score displays an additional peak around 1702 cm^{-1}. These observations suggest that both linear and logarithmic regression can be used, although different results should be expected. The study of such differences can be instrumental to define which type of relation exists, if any, between the stressing conditions imposed in the experiments and the metabolomic output, bearing in mind that different spectral regions correspond to diverse molecules and is therefore expectable that not all metabolits respond in the same way to the same stress.

Although the score plot is a synopsis of the other two plots, a careful inspection of all the three windows yielded by the *mms* function is advisable in order to get the maximum possible information from each experiment.

By averaging the replicas, a 6×1194 matrix was obtained and subject to *mms* again, obtaining the result depicted in

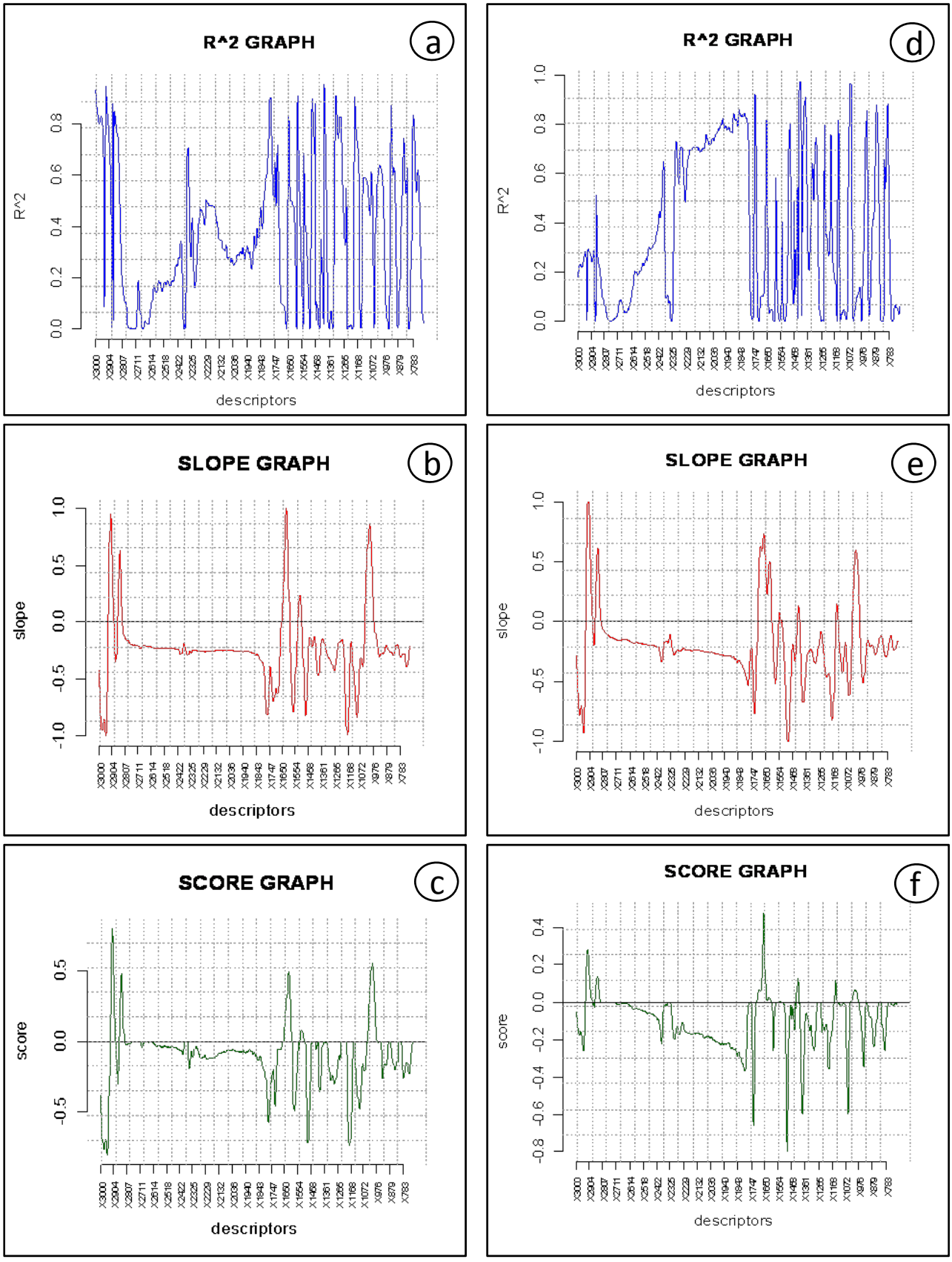

Fig. (2). Response diagram of three independent repetitions. (**a**) R^2 linear response diagram, (**b**) slope diagram of linear regression, (**c**) score diagram obtained from linear regression. (**d**), (**e**) and (**f**) are logarithmic response diagrams corresponding to (**a**), (**b**) and (**c**).

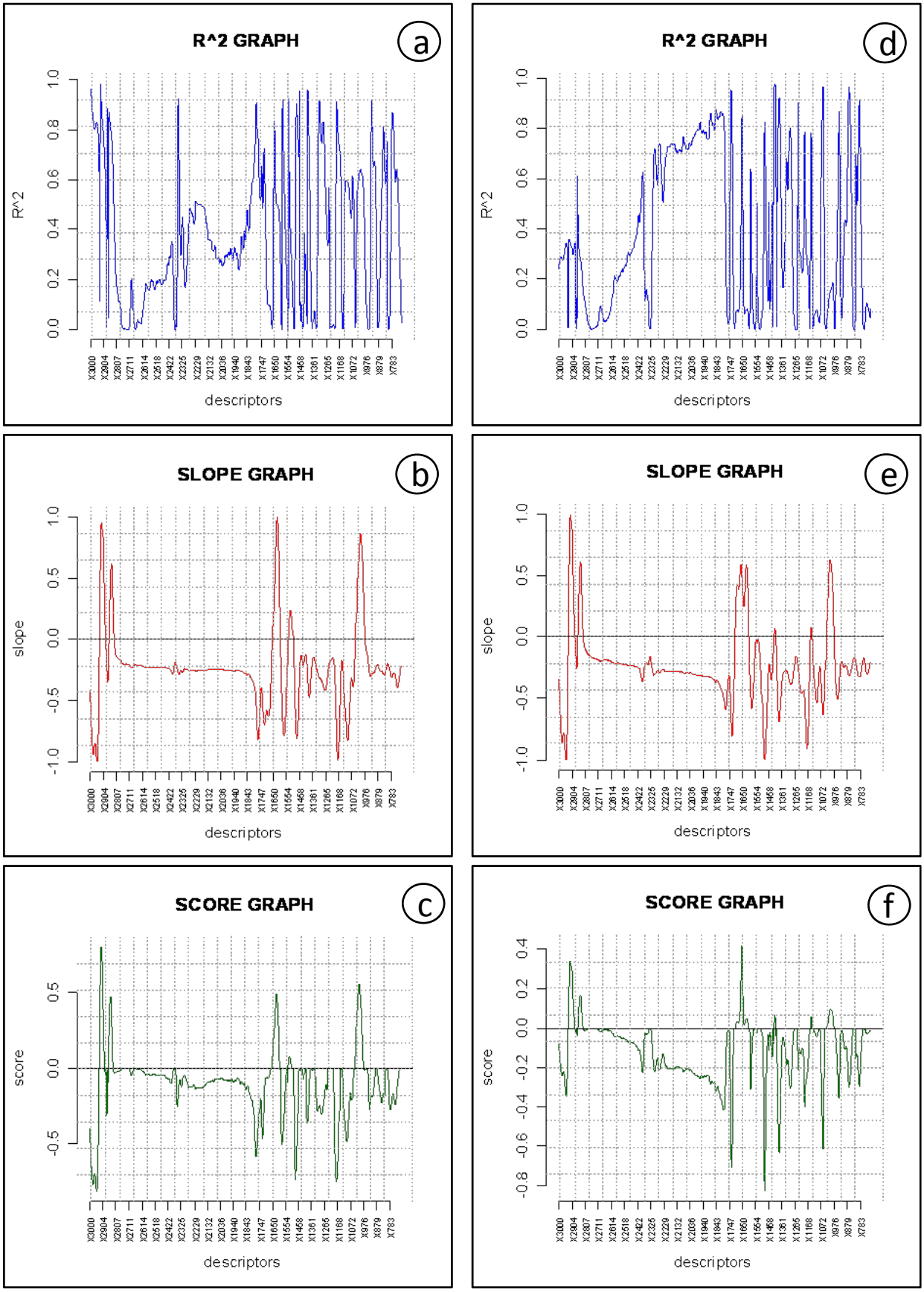

Fig. (3). Response diagram of average data. (**a**) R^2 linear response diagram, **b**) slope diagram of linear regression, (**c**) score diagram obtained from linear regression. (**d**), (**e**) and **(f)** are logarithmic response diagrams corresponding to (**a**), (**b**) and (**c**).

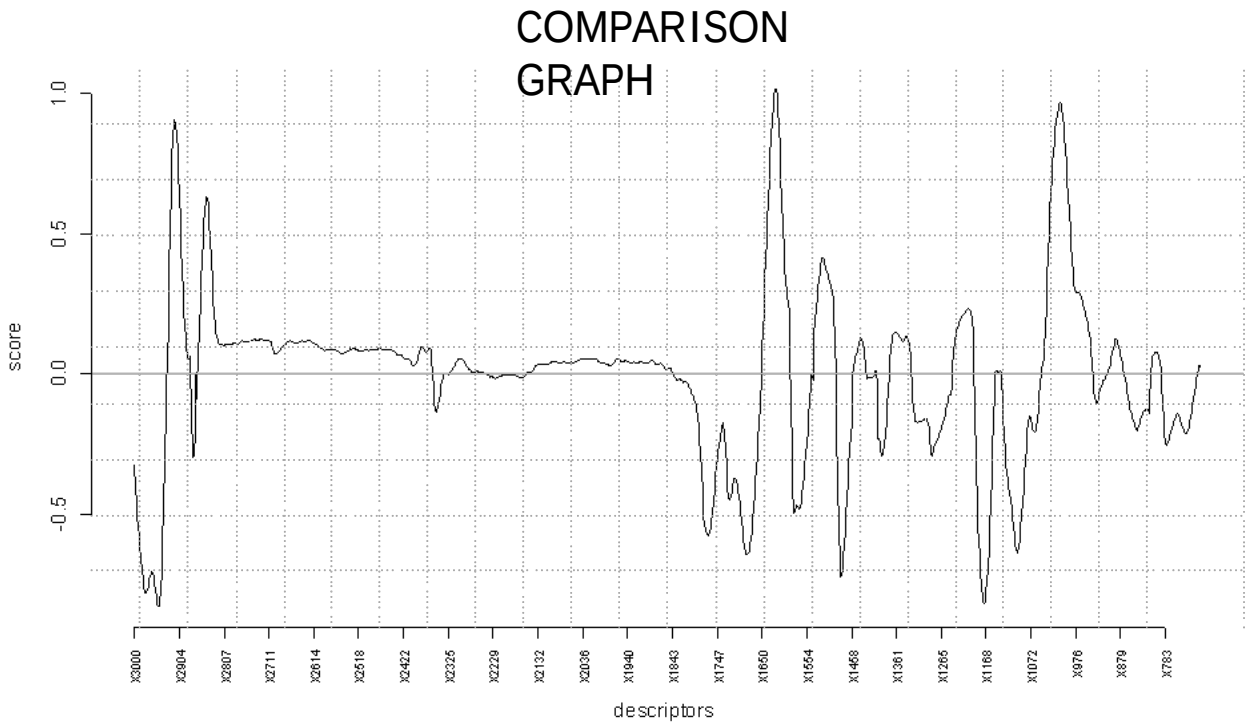

Fig. (4). Difference graph of yeast cells subject to 200 mg/l SO_2.

Fig. (**3**). A comparison of the six panels of Fig. (**2**) with the corresponding of Fig. (**3**) shows little if any difference, mainly in the magnitude of the value reported in the ordinate. This observation indicates that no significant difference is generated by averaging the original spectral values, when replicated spectra are quite similar. The fact that averaging values reduces the objects should not be considered a crucial point in terms of calculation time, in fact with an Intel Core Duo T5500 (1.66 MHz) processor the whole *mms* procedure took 4.5 sec when the matrix with all replicas was processed and 3.5 sec when the averaged values were analyzed.

The evaluation of the score plot should consider that in our case derivatives and not spectral intensities have been processed, meaning that positive or negative score peaks do not signify an increase or a decrease of the vibrations, but rather that the response was observed on the ascending or descending side of a spectral peak, respectively. A tentative evaluation of the score plots (Fig. **2c**,**f**) revealed that one of the most prominent peaks was located in the fatty acids region 2950-2850 cm^{-1}, showing a more intense response with the linear rather than with the logarithmic regression analysis. Other significant responses were present in the lipid-esters and RNA-DNA region (1741-1708 cm^{-1}), in the proteins 1695-1675 cm^{-1}, and amide I peak (1655-1637 cm^{-1}) area. Proteins and lipids (1515-1457 cm^{-1}), RNA-DNA backbones (1160-1076 cm^{-1}) and the carbohydrates region 1000 cm^{-1} were other locations with significant responses. Altogether, the SO_2 challenge seemed to modify several molecules, among which the lipids, which could be the primary targets of this compound, given their localization in the cell membrane. Considering that our cells were not able to reproduce and therefore to synthesize DNA, the DNA/RNA reactivity should have been caused only by translation, meaning that the cells reacted immediately with the activation of some metabolic pathway. Particularly interesting is marked response in the region of the aldehydes (1740-1725 cm^{-1}), since the most known reaction of *S. cerevisiae* subject to SO_2 is an increase in acetaldehyde, which binds covalently to the sulfur dioxide. This finding is a preliminary confirm that indeed MMS can help individuating the molecules reacting to a given stress.

The *compare* function was carried out to compare the *mms* score plots generated by two identical experiments in which the only difference was the concentration of the sulfur dioxide (200 *vs* 400 mg/l). the resulting comparison plot (Fig. **5**) shows that some areas are differently affected by the two stressing conditions, suggesting that maybe diverse metabolisms are employed in the two cases, although more accurate investigations are compulsory to support this supposition.

In general, the *compare* function has not been designed to immediately draw conclusions on differential responses, but rather to highlight these differences and to let the investigator focus on the spectral regions involved in these discrepancies.

The *diff* function is a simple algorithm that calculates the mean spectral difference between resting and stressed cells, with outputs ranging from $-\infty$ to $+\infty$.

This function does not require a reference vector as *mms*, but works only and directly with spectral data. The fact that the difference plot (Fig. **4**) has little resemblance with the graphs of Figs. (**2**,**3**) is not unexpected because the rationale of the two functions, *mms vs dif*, is radically different: the

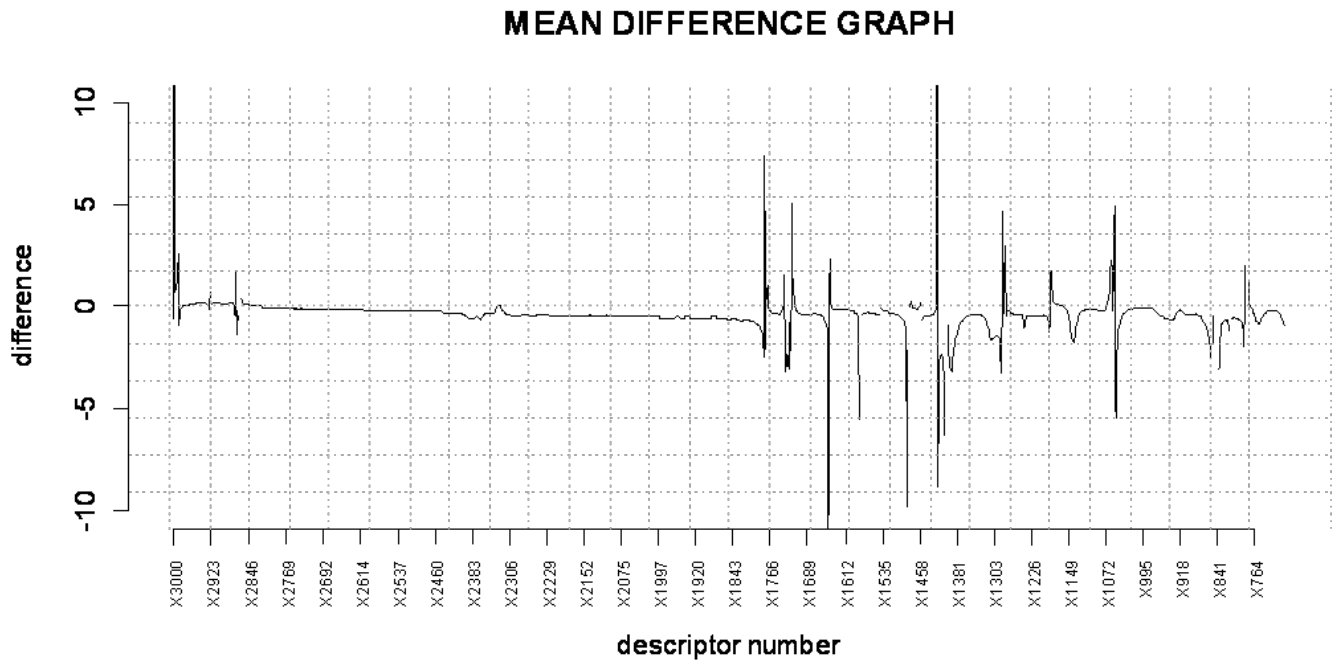

Fig. (5). Comparison between score results from MMS analyses of yeast cells subject to 200 and 400 mg/l SO_2.

former investigates where a linear or logarithmic relation exists between experimental conditions and metabolomic response, the latter simply calculates in which parts of the spectrum the stress has caused more differences. In those areas where there is a great degree of difference but little *score* the stress has caused large metabolic changes, but these are not apparently related to the stressing conditions. Whether such a spectral area can be of interest for further analyses or elaborations should be decided by the investigator; the software is designed only to highlight the phenomenon.

CONCLUSIONS

Metabolomic data from stress response experiments are typically analyzed by multivariate statistical analysis often producing the so called "metabolomic fingerprint" [20-21]. MMS has not been designed as an alternative to these consolidated approaches, but rather as an additional tool to help researchers in pinpointing the regions with a more marked response. After this preliminary analysis, further investigations will be necessary to elucidate the metabolic and physiological meaning of the cell reaction to the tested challenge. Moreover, MMS score plot can be conveniently interpreted as a "response plot", which is likely to be specific for each stress, according to our preliminary data (not shown). Further investigations can elucidate the level of specificity of these response plots, whether some signals are a sort of general stress response whereas others are specific.

This article has shown the potentialities of the package using matabolomic data obtained with the FTIR technology. Results from any other metabolomic technique such as, HPLC, CE, GC, NMR, MS/MS etc [22] can be conveniently analyzed with this software. This articles deals mainly with the package graphic outputs for reasons of space and clarity; obviously, the investigator will take advantage of the tables resulting from all analyses and of the ease to export them for further use with any other suitable software.

REFERENCES

[1] M. Kummerle, S. Scherer and H. Seiler, "Rapid and reliable identification of food-borne yeasts by Fourier-transform infrared spectroscopy", *Appl. Environ. Microbiol.*, vol. 64, pp. 2207-2214, June 1998.

[2] E. M. Timmins, S. A. Howell, B. K. Alsberg, W. C. Noble, and R. Goodacre, "Rapid differentiation of closely related Candida species and strains by pyrolysis-mass spectrometry and Fourier transform-infrared spectroscopy". *J. Clin. Microbiol.*, vol. 36, pp. 367-374, February 1998.

[3] K. Tintelnot, G. Haase, M. Seibold, *et al.*, "Evaluation of phenotypic markers for selection and identification of *Candida dublinien-sis*", *J. Clin. Microbiol.*, vol. 38, pp. 1599-1608, April 2000.

[4] M. Wenning, H. Seiler and S. Scherer, "Fourier-transform infrared microspectroscopy, a novel and rapid tool for identification of yeasts", *Appl. Environ. Microbiol.*, vol. 68, pp. 4717-4721, October 2002.

[5] A. Oust, T. Moretro, C. Kirschner, J. A. Narvhus and A. Kohler, "FT-IR spectroscopy for identification of closely related lactobacilli", *J. Microbiol. Methods.*, vol. 59, pp. 149-162, November 2004.

[6] M. Essendoubi, D. Toubas, M. Bouzaggou, J. M. Pinon, M. Manfait and G. D. Sockalingum, "Rapid identification of *Candida* species by FT-IR microspectroscopy", *Biochim. Biophys. Acta.*, vol. 1724, pp. 239-247, August 2005.

[7] D. Naumann, D. Helm and H. Labischinski, "Microbiological characterizations by FT-IR spectroscopy", *Nature*, vol. 351, pp. 81-82, May 1991.

[8] G. D. Sockalingum, W. Bouhedja, P. Pina, P. Allouch, C. Bloy and M. Manfait, "FT-IR spectroscopy as an emerging method for rapid characterization of microorganisms", *Cell. Mol. Biol. (Noisy-le-grand)*, vol. 44, pp. 261-269, February 1998.

[9] C. Yu and J. Irudayaraj, "Spectroscopic characterization of microorganisms by Fourier transform infrared microspectroscopy", *Biopolymers.*, vol. 77, pp. 368-377, April 2005.

[10] M. Mecozzi, F. Onorati, F. Oteri and A. Sarni, "Characterization of a bioassay using the marine alga *Dunaliella tertiolecta* associated with spectroscopic (visible and infrared) detection", *Int. J. Environ. Pollut.,* vol. 32, pp. 104-120, January 2008.

[11] H. Zhao, Y. Kassama, M. Young, D. B. Kell and R. Goodacre, "Differentiation of *Micromonospora* isolates from a coastal sediment in Wales on the basis of Fourier transform infrared spectroscopy, 16S rRNA sequence analysis, and the amplified fragment length polymorphism technique", *Appl. Environ. Microbiol.*, vol. 70, pp. 6619-6627, November 2004.

[12] W. Zeroual, M. Manfait and C. Choisy, "FT-IR spectroscopy study of perturbations induced by antibiotic on bacteria (*Escherichia coli*)", *Pathol. Biol. (Paris)*, vol. 43, pp. 300-305, April 1995.

[13] E. Breierova, Z. Hromadkova, E. Stratilova, V. Sasinkova and A. Ebringerova, "Effect of salt stress on the production and properties of extracellular polysaccharides produced by *Cryptococcus laurentii*", *Z. Naturforsch. [C]*, vol. 60, pp. 444-450, May-June 2005.

[14] K. Maquelin, C. Kirschner, L. P. Choo-Smith, N. van den Braak, H. P. Endtz, D. Naumann and G. J. Puppels, "Identification of medically relevant microorganisms by vibrational spectroscopy", *J. Microbiol. Methods.*, vol. 51, pp. 255-271, November 2002.

[15] K. Lemmer, D. Naumann, B. Raddatz and K. Tintelnot, "Molecular typing of *Cryptococcus neoformans* by PCR fingerprinting, in comparison with serotyping and Fourier transform infrared-spectroscopy-based phenotyping", *Med. Mycol.*, vol. 42, pp. 135-147, April 2004.

[16] G. Fischer, S. Braun, R. Thissen and W. Dott, "FT-IR spectroscopy as a tool for rapid identification and intra-species characterization of airborne filamentous fungi", *J. Microbiol. Methods*, vol. 64, pp. 63-77, January 2006.

[17] C. A. Rebuffo-Scheer, C. Kirschner, M. Staemmler and D. Naumann, "Rapid species and strain differentiation of non-tubercoulous mycobacteria by Fourier-Transform Infrared microspectroscopy", *J. Microbiol. Methods.*, vol. 68, pp. 282-290, February 2007.

[18] M. Beekes, P. Lasch and D. Naumann, "Analytical applications of Fourier transform-infrared (FT-IR) spectroscopy in microbiology and prion research", *Vet. Microbiol.*, vol. 123, pp. 305-319, August 2007.

[19] B. Moen, A. Oust, O. Langsrud, N. Dorrell, G. L. Marsden, J. Hinds, A. Kohler, B. W. Wren and K. Rudi, "Explorative multifactor approach for investigating global survival mechanisms of *Campylobacter jejuni* under environmental conditions", *Appl. Environ. Microbiol.*, vol. 71, pp. 2086-2094, April 2005.

[20] O. Fiehn, "Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks", *Comp. Funct. Genomics,* vol. 2, pp. 155-168, April 2001.

[21] J. Coates, *Interpretation of Infrared Spectra, A practical Approach. Encyclopedia of Analytical Chemistry.* John Wiley & Sons Ltd, Chichester, 2000.

[22] S. Vaidyanathan, G. G. Harrigan, R. Goodacre, *Metabolome Analyses: Strategies for Systems Biology*. Springer, 2005.

Received: February 14, 2008 Revised: March 10, 2008 Accepted: March 14, 2008