Available online at www.sciencedirect.com
SCIENCE DIRECT®
PERGAMON
Phytochemistry 62 (2003) 919–928
PHYTOCHEMISTRY
www.elsevier.com/locate/phytochem

Metabolic fingerprinting of salt-stressed tomatoes

Helen E. Johnson, David Broadhurst, Royston Goodacre, Aileen R. Smith*

Institute of Biological Sciences, Cledwyn Building, University of Wales, Aberystwyth, Ceredigion, SY23 3DD, Wales, UK

Received 20 September 2002; received in revised form 5 November 2002

Abstract

The aim of this study was to adopt the approach of metabolic fingerprinting through the use of Fourier transform infrared (FT-IR) spectroscopy and chemometrics to study the effect of salinity on tomato fruit. Two varieties of tomato were studied, Edkawy and Simge F1. Salinity treatment significantly reduced the relative growth rate of Simge F1 but had no significant effect on that of Edkawy. In both tomato varieties salt-treatment significantly reduced mean fruit fresh weight and size class but had no significant affect on total fruit number. Marketable yield was however reduced in both varieties due to the occurrence of blossom end rot in response to salinity. Whole fruit flesh extracts from control and salt-grown tomatoes were analysed using FT-IR spectroscopy. Each sample spectrum contained 882 variables, absorbance values at different wavenumbers, making visual analysis difficult and therefore machine learning methods were applied. The unsupervised clustering method, principal component analysis (PCA) showed no discrimination between the control and salt-treated fruit for either variety. The supervised method, discriminant function analysis (DFA) was able to classify control and salt-treated fruit in both varieties. Genetic algorithms (GA) were applied to identify discriminatory regions within the FT-IR spectra important for fruit classification. The GA models were able to classify control and salt-treated fruit with a typical error, when classifying the whole data set, of 9% in Edkawy and 5% in Simge F1. Key regions were identified within the spectra corresponding to nitrile containing compounds and amino radicals. The application of GA enabled the identification of functional groups of potential importance in relation to the response of tomato to salinity.

Keywords: *Lycopersicon esculentum*; Tomato; Fourier transform infrared spectroscopy; Salt stress; Metabolic fingerprinting

1. Introduction

The increased salinization of agricultural land, due to human activities, agricultural practises and natural processes, is having a negative impact on soil fertility and crop productivity in many parts of the world. Recent reports published by the Food and Agriculture Organization (FAO) of the United Nations state that of the 260 million hectares of irrigated land world wide, 80 million are affected to some extent by salinization (FAO, 2002a). In an earlier publication (FAO, 2002b) it was estimated that salinization was reducing the world's irrigated area for crop production by 1–2% every year. The problem is most severe in arid and semi-arid regions of the world (Epstein et al., 1980) where the use of irrigation practises is extensive and necessary to overcome the problems of drought, intermittent rain and to extend the crop-growing season (Toenniessen, 1984; Ellis and Mellor, 1995). In these regions the problem is exacerbated due to the low rainfall, high temperatures and high irradiance levels. Farmers, particularly in these parts of the world, are changing growing practises to cope with the economic effects of salinization. In Cape Verde, farmers have tripled horticultural production (in terms of tonnage) from 1991 to 1999 by changing from growing sugar cane to high-value horticultural crops such as tomatoes (FAO, 2002b).

Tomato is one of the major horticultural cash crops in the world (Hilhorst et al., 1998). However, increasing salinity levels negatively effect germination, plant growth and fruit yield as described in a review by Cuartero and Fernandez-Munoz (1999). Salinization also results in the increased occurrence of the physiological disorder blossom end rot (BER) in many tomato varieties (Brown and Ho, 1993; Ho et al., 1993). The occurrence of BER is related to a decrease in the absorption and translocation of calcium ions to the fruit (Franco et al., 1994) and significantly reduces the marketable value

* Corresponding author. Tel.: +44-1970-622343; fax: +44-1970-622307.
E-mail address: ars@aber.ac.uk (A.R. Smith).

0031-9422/03/$ - see front matter © 2003 Elsevier Science Ltd. All rights reserved.
doi:10.1016/S0031-9422(02)00722-7

of the crop. The mechanisms imparting salt tolerance are complex and vary greatly depending on the level and duration of the salt stress, the species and the ontogeny of the plant. This is reflected in the wealth of scientific publications available, highlighted in a review by Flowers and Yeo (1995). As many mechanisms and adaptive processes are inter-linked the study of a single mechanism or process in isolation often oversimplifies the problem.

The ability to obtain biochemical fingerprints could be of both scientific and commercial importance in studies on for example the characterization of genetic mutants, plant breeding and plant responses to environmental stresses. Hence, the approach of metabolic fingerprinting was applied to study the biochemistry of tomato fruits grown in saline conditions and to study the metabolic discrimination between varieties. Tomato was selected as a model species for these studies due to its commercial importance as a horticultural crop. Two tomato varieties were used in this work, Edkawy, an Egyptian beefsteak tomato variety previously reported to have an increased salt-tolerance (Mahmoud et al., 1986a, b; Cuartero et al., 1992) and Simge F1.

Metabolic fingerprinting is the rapid classification of samples according to their origin or biological provenance (Fiehn, 2001) where it is not initially necessary or feasible to determine the levels of metabolites individually but to develop a high-throughput technique which enables a snap-shot of the metabolic composition at a given time. One approach has been the use of gas chromatography mass spectrometry to obtain metabolic profiles of plant tissues (Fiehn, 2001). Alternatively, Fourier transform infrared spectroscopy (FT-IR) has been used successfully in the differentiation of closely related microorganisms (Timmins et al., 1998) and the identification of biomarkers in *Bacillus* spores (Goodacre et al., 2000).

FT-IR spectroscopy presents itself as an ideal candidate for high-throughput metabolic fingerprinting. FT-IR is a physico-chemical method that measures predominantly the vibrations of bonds within functional groups (Griffiths and de Haseth, 1986) and generates a spectrum that can be regarded as a biochemical or metabolic 'fingerprint' of the sample. However FT-IR spectra are complex, containing 882 variables per sample making visual analysis very difficult. Hence, a range of chemometric and data mining techniques are applied to analyse these multivariate data. Two different chemometric methods were applied here, principal component analysis (PCA) and discriminant function analysis (DFA). Genetic algorithms (GAs), an evolutionary computational method, were applied as a variable selection method prior to discriminant multiple linear regression (D-MLR) analysis to deconvolve these hyper-spectral data sets in terms of the important discriminatory regions within the spectra.

The objective of this research was to adopt the approach of metabolic fingerprinting using FT-IR spectroscopy and chemometrics, to distinguish between control and salt-treated fruit and to investigate the use of GA-D-MLR for identifying discriminatory biomarkers for susceptible and salt tolerant tomato varieties.

2. Results and discussion

2.1. Growth and fruit yield

In order to establish the degree of salt tolerance possessed by each variety screening experiments were performed and results were analysed using classical growth analysis. The plants were grown in a hydroponic drip irrigation system that enabled two independent treatments, a control and a saline treatment of 0.4% w/v NaCl. Table 1 shows the effect of 0.4% w/v NaCl on the two rate parameters, relative growth rate (RGR) and net assimilation rate (NAR), in both Edkawy and Simge F1. RGR is a measure of overall plant efficiency whereas NAR is an assessment of the photosynthetic efficiency (Hunt, 1982). The increased salinity level had no significant effect on the RGR or NAR in Edkawy confirming the previously reported salt tolerance (Mahmoud et al., 1986a, b; Cuartero et al., 1992). However, the RGR of Simge F1 was significantly reduced by salinity on both the east and west facing sides of the greenhouse. Despite the significant decrease in the RGR of Simge F1 in response to salinity no significant difference in the NAR was observed. Due to the orientation of the greenhouse, the west facing side typically received a higher irradiance level after midday compared to the east facing side. This east/west variation was incorporated into the metabolic fingerprinting as fruit were selected from plants grown on both sides of the greenhouse.

Defining susceptibility and tolerance to salinity is difficult as there is often no clear cut-off point between these two attributes. The response of a plant to salinity

Table 1
A comparison of the effect of 0.4% w/v NaCl on the relative growth rate (RGR) and net assimilation rate (NAR) of Edkawy and Simge F1 tomato varieties when grown in a hydroponic drip irrigation system

	East facing			West facing		
	Control	Saline		Control	Saline	
Edkawy						
RGR day^{-1}	0.1327	0.0987	ns	0.1435	0.1290	ns
NAR g day^{-1}	5.2×10^{-4}	3.7×10^{-4}	ns	6.9×10^{-4}	6.3×10^{-4}	ns
Simge F1						
RGR d^{-1}	0.1405	0.1102	5%	0.1321	0.0972	0.1%
NAR g day^{-1}	5.8×10^{-4}	4.8×10^{-4}	ns	5.3×10^{-4}	4.6×10^{-4}	ns

Data collected 14 days after initial salt application. $N=5$. Data were analysed using Two-way analysis of variance. ns = on significant; 5 and 0.1% indicate probability of significance.

is affected by many different factors including growth stage, magnitude and duration of the stress and climatic conditions. In this experiment the effect of salinity on vegetative growth was monitored over a short time period and although there are some similarities between the growth parameters for each variety, the significant decrease in RGR in Simge F1 indicated an increased susceptibility to salinity compared to Edkawy. Although it is important to assess the effect of salinity on vegetative growth, ultimately it is the effect of increased salinity on fruit yield and quality that is of economic importance.

Table 2 compares the fruit yield and BER occurrence in both Edkawy and Simge F1 under control and saline conditions. All fruit were harvested from the 1st and 2nd trusses over a 26 day period in the case of Edkawy and a 32 day period for Simge. Due to the large size of the Edkawy fruit fewer fruit grew per truss, hence the harvest period was shorter before all ripe fruit were harvested.

Salinity had no significant effect on the number of fruit harvested in either Edkawy or Simge F1 (Table 2), with 43 and 46% of the total fruit harvested from salt-treated plants in Edkawy and Simge F1, respectively. However, marked differences can be seen between the two varieties and treatments when comparing the effect of salinity on the average fruit weight and size class (Table 2). Salinity resulted in a decrease in fresh weight in both Edkawy and Simge F1 (Table 2). However, the magnitude of the decrease was much greater in Simge F1 than in Edkawy. In Simge F1 the mean fruit fresh weight was reduced by 54% in response to salinity, compared to a 28% reduction in Edkawy. The reduction in mean fruit size class (Table 2) was relatively small in Edkawy compared to the marked decrease in Simge F1 in accord with the previous assessments of salt tolerance.

The most noticeable effect of salinity on the fruit of Edkawy and Simge was the increased occurrence of BER. Despite the reported salt tolerance of Edkawy these data show an increased susceptibility to BER comparable to that observed in Simge F1 (Table 2). This increased occurrence of BER significantly reduced the marketable yield of both crops.

Table 2
The effect of 0.4% w/v NaCl on the fruit yield and blossom end rot occurrence in two tomato varieties Edkawy and Simge F1

	Edkawy		Simge F1	
	Control	Saline	Control	Saline
% Total yield	57	43	54	46
Mean fresh weight, g	151.4	109.7	126.6	58.4
Mean size class	5.1	4.4	5.5	2.7
% BER		50		34

2.2. Cluster analyses of FT-IR spectra

The tissue extracts of fully ripe fruit samples were analysed using FT-IR. It should be noted that the two varieties (control and salt-treated) were analysed individually as separate experiments. Fig. 1 shows an example FT-IR spectrum for each variety. These FT-IR spectra and all the others collected look very similar, they all show broad and complex contours and it is difficult to identify key features by eye. Such spectra readily illustrate the need to employ chemometric techniques for their analysis.

The characteristic CO_2 peaks at wavenumbers 656–683 cm^{-1} and 2272–2403 cm^{-1} were removed and replaced by a trend line prior to further analysis. Fig. 2a and b shows PCA plots for Edkawy and Simge F1, respectively, based on the FT-IR spectra fingerprints of each. It can be seen in these figures that there is no discrimination between the control and salt-treated fruit, indicated as 0 and 1 respectively, for either variety. It can also been seen that there is a large intrinsic variability within the data sets, 80% of which is accounted for by the first principal component (PC).

DFA models for Edkawy and Simge F1 were then produced using knowledge of the treatment group structure of each data set, hence termed a supervised technique. The number of PCs used in each DFA model was optimised by cross-validation, using a training and test data set (see Experimental 4.6). The resultant plots are shown in Figs. 3 and 4, corresponding to Edkawy and Simge F1, respectively. Clearly, the DFA model was able to discriminate between the control and salt-treated fruits in both varieties. However, the separation was more distinct between the control and salt-treated fruits of Simge F1 than between those of Edkawy, with only one control fruit misclassified in the test set. This trend may be due to a greater biochemical difference between the control and salt-treated fruit of Simge F1 as

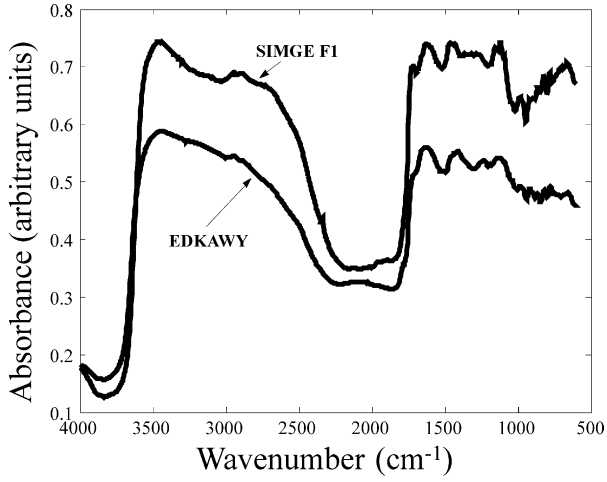

Fig. 1. Representative FT-IR spectra from whole tomato fruit flesh of Edkawy and Simge F1.

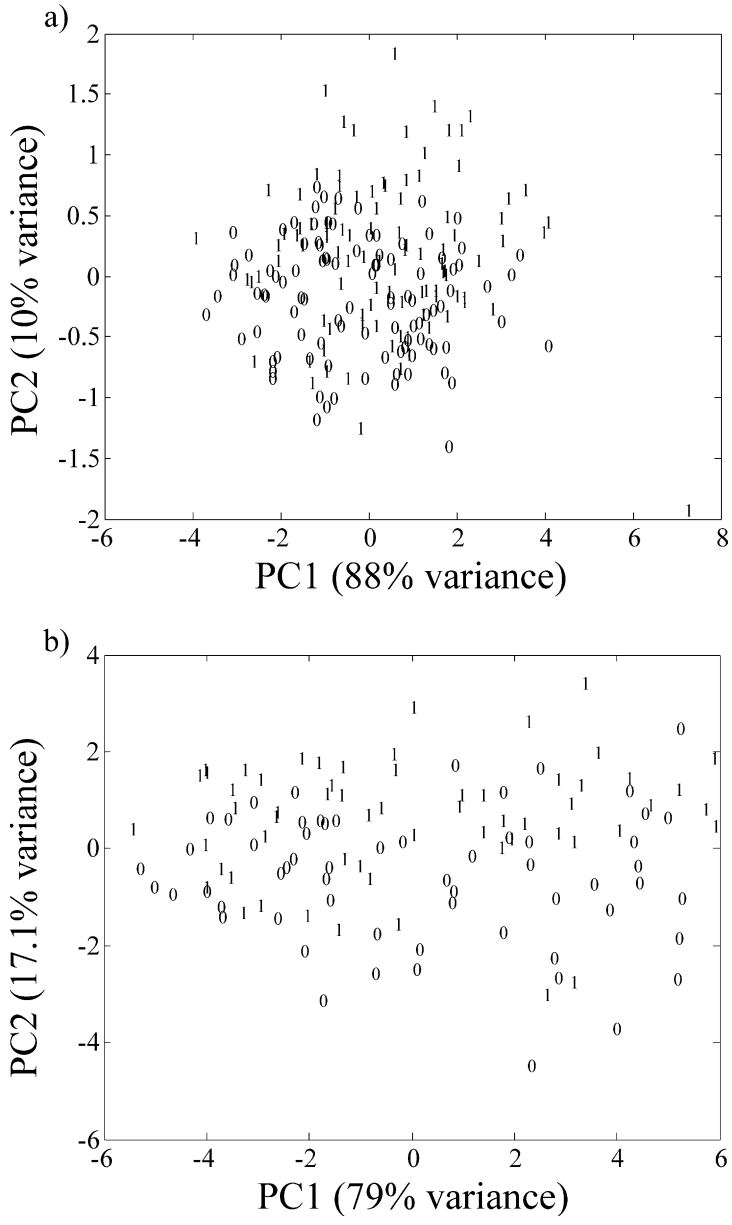

Fig. 2. Principal component analysis (PCA) models for (a) Edkawy and (b) Simge F1 showing no discrimation between the control fruit (0) and the fruit from salt-treated plants (1) in either variety. (a) Edkawy (b) Simge F1.

a result of an increased sensitivity to salinity in comparison to Edkawy.

Classification of fruit according to treatment by DFA indicated that the FT-IR spectral fingerprints contained discriminatory biochemical information. However, it was very difficult to ascertain from the DFA model which parts of the spectra were important for discriminating between the fruit samples. The first discriminant function (DF1) loadings were studied (data not shown) but this was inconclusive as no obvious regions were identified as being of key importance for discriminating between control and salt-grown fruits. Therefore the data mining technique of genetic algorithms (GAs) was employed.

2.3. Genetic algorithms

GAs have previously been applied for the identification of discriminatory variables within multivariate data sets (Broadhurst et al., 1997; Yoshida et al., 2001; Ellis et al., 2002). The GA can be run many times on a single data set producing multiple 'optimal' models each using a subset of the available data (the user can predefine the size of each subset). The variables used in each model can then be aggregated and the frequency of variable selection can be calculated. In this work the program was run 50 times and used five variables for both tomato varieties, generating 50 independent models for each. The number of variables was set to five, as this

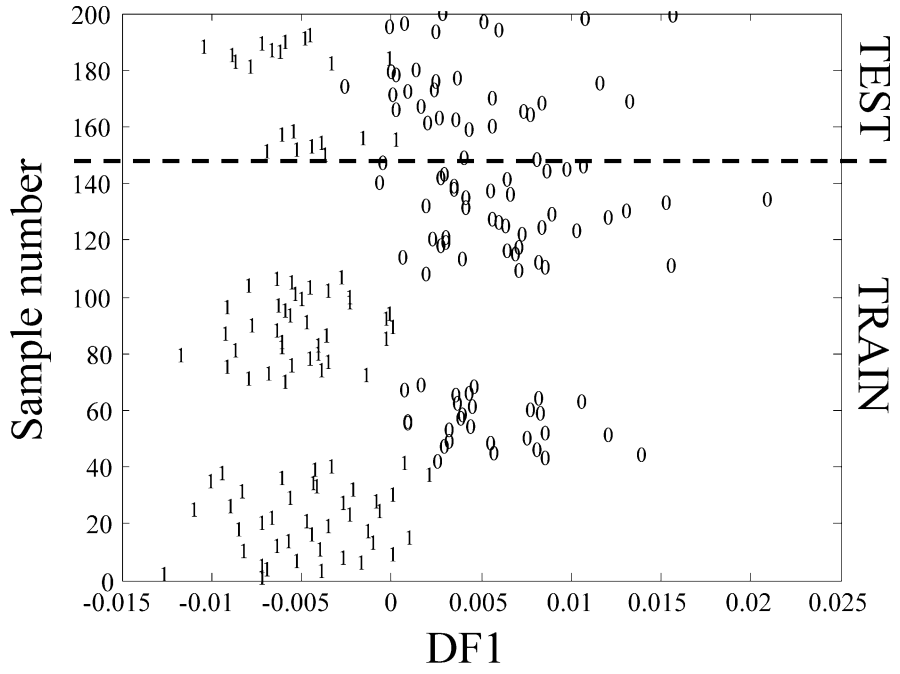

Fig. 3. Discriminant function analysis (DFA) model using 20 principal components (PCs) accounting for 99.99% total variance derived from the raw FT-IR spectral data for Edkawy tomatoes. Training data set contained samples 1 to 149 and test data set contained samples 150–200. The number of PCs to be used for DFA was optimised using the training data set and then the test data were projected onto the DFA model. The model shows discrimination between control and salt-treated Edkawy tomato fruit although there are misclassified samples in both the training and test data sets. 0 = control fruit and 1 = salt treated fruit.

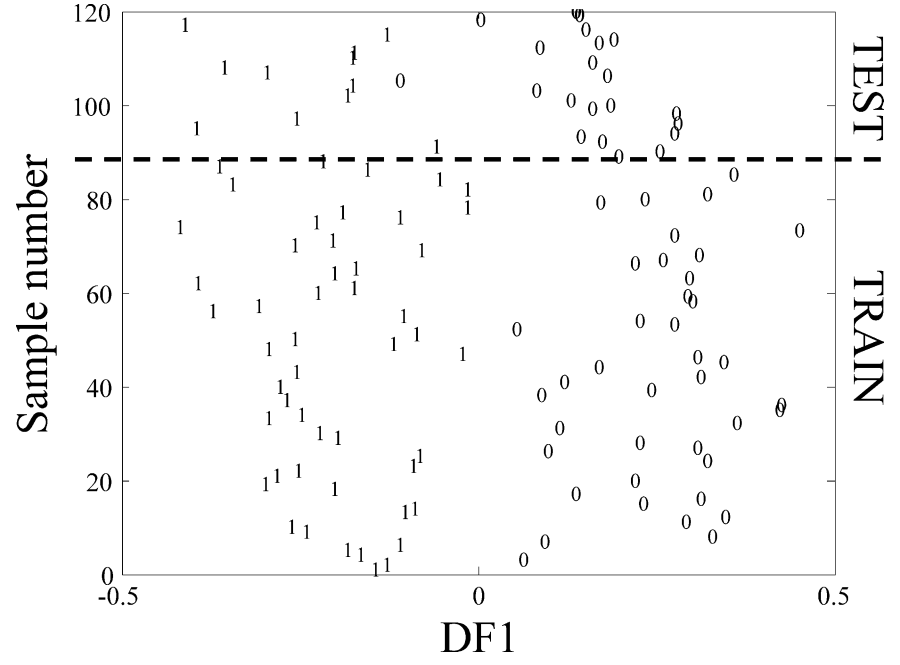

Fig. 4. Discriminant function analysis (DFA) model using 20 principal components (PCs) accounting for 99.99% total variance derived from the raw FT-IR spectral data for Simge F1 tomatoes. Training data set contained samples 1–90 and test data set contained samples 91–120. The number of PCs to be used for DFA was optimised using the training data set and then the test data were projected onto the DFA model. The model shows discrimination between control and salt-treated Simge F1 tomato fruit. 0 = control and 1 = salt treated.

was the minimum number of variables required to obtain adequate discrimination between the sample groups. The aim was to identify key regions of importance within the spectra for discrimination between control and salt-treated tomatoes. The results obtained are shown in Figs. 5 and 6 corresponding to Edkawy and Simge F1, respectively. On average the total error of classification for the models was 9% in Edkawy and 5% in Simge F1. Given the complexity of the data and the extent to which dimensionality of the data set was reduced, errors of below 10% reflect the power of GA. The errors recorded for the GA rules were comparable to those observed in the DFA plots (Figs. 3 and 4).

The GA used similar regions within the spectra for the discrimination between control and salt-treated tomato fruit in each variety. The main regions used by

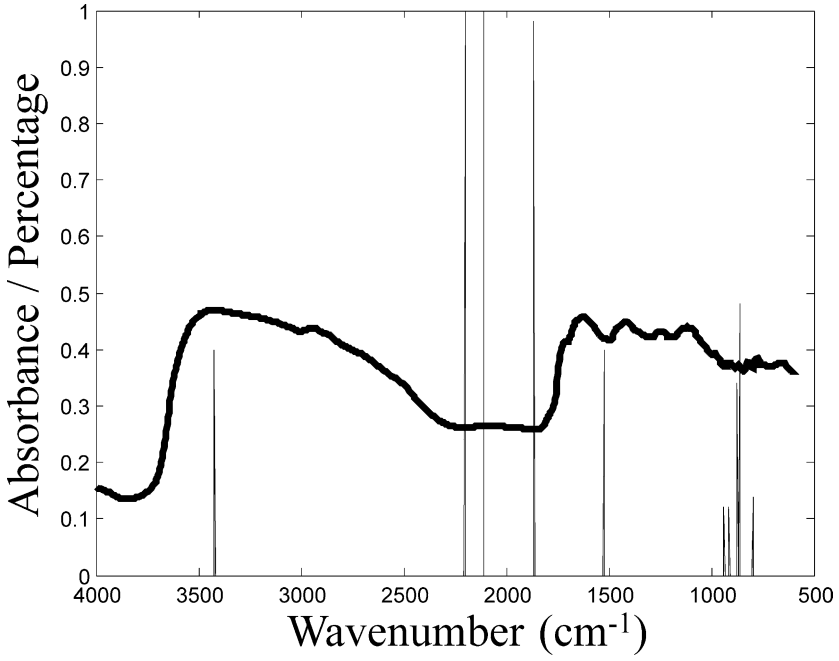

Fig. 5. Variable selection percentage by 50 independent genetic algorithm (GA) models, with each model using only 5 variables, for the discrimination between control and salt-treated Edkawy tomato fruit samples based on FT-IR spectral data.

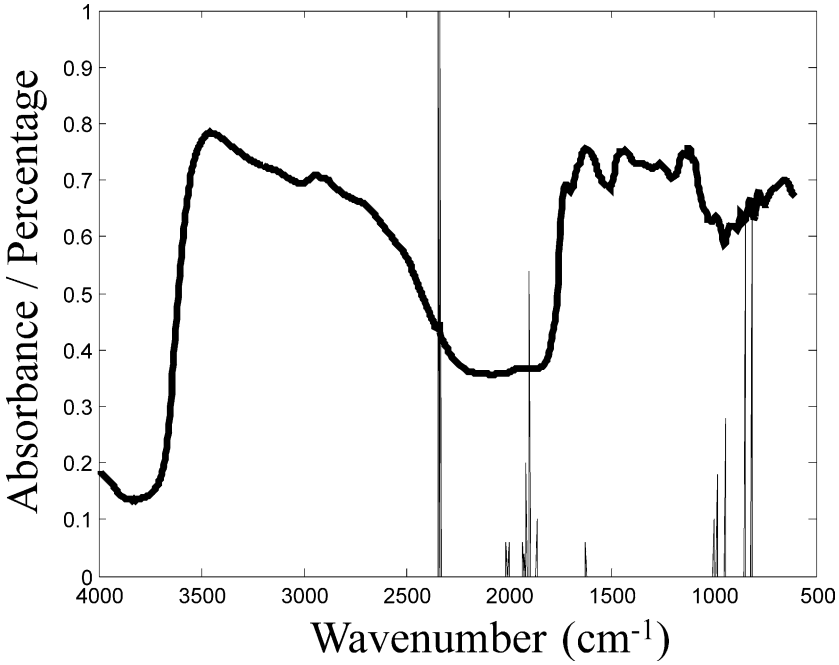

Fig. 6. Variable selection percentage by 50 independent genetic algorithm (GA) models, with each model using only 5 variables, for the discrimination between control and salt-treated Simge F1 tomato fruit samples based on FT-IR spectral data.

the GA models for the classification of Edkawy fruit were around wavenumbers 2200, 2110 and 1860 cm^{-1}. A region within this wavenumber range, from 2270 to 1960 cm^{-1}, was also identified as being key for the discrimination between Edkawy fruit samples when analysing these data using another evolutionary method, genetic programming (GP) (Johnson et al., 2000). The selection of this region by two independent modelling methods strengthens the argument that this region must be of importance for the classification of control and salt-treated tomato fruit. The GA models for Simge F1 consistently used an area just outside this key region, around 2300 cm^{-1} and a second region of variables, around 1880 cm^{-1}, an area also used by the GA models for Edkawy. The GA also selected the region 800–900 cm^{-1} in the classification of both tomato varieties. Although the biochemical significance of these regions requires further investigation the application of GA has enabled the following interpretation to be made. The region within the IR spectrum spanning from 2100 to 2300 cm^{-1} corresponds to saturated and unsaturated nitrile compounds and C≡N stretch. The region towards the end of the spectra (800–900 cm^{-1}) contains a strong broad peak for NH_2 (an amino radical) and other N compounds according to IRmentor software version 2.0.0.17 (BarSpec Systems Inc.) and Degen (1997). Collectively these spectral regions selected by the GA for discrimination between control and salt-treated tomatoes are indicating a shift in the biochemistry of nitrile containing compounds although future mass spectrometric studies would be required to confirm this.

3. Conclusion

This inductive reasoning approach of applying FT-IR spectroscopy coupled with chemometric techniques enabled us to classify control and salt-treated tomato fruits with respect to their metabolic fingerprints. The application of GA identified key regions within the spectra for the discrimination between the control and salt-treated fruit of each variety. These analyses suggest specific functional groups of importance in relation to tomato responses to salinity, which may not have emerged as key compounds through conventional biochemical analysis. The logical extension of this research is to adopt a metabolite profiling approach (Fiehn, 2001) targeting specific compound groups. The results from the GAs indicated that future research should target saturated and unsaturated nitrile compounds and cyanide containing compounds. These results are in agreement with the hypothesis proposed in Johnson et al. (2000), that the increased occurrence of cyanide or nitrile containing compounds might be attributed to the detoxification of hydrogen cyanide. Although speculative, one possible interpretation is that hydrogen cyanide is a by-product produced during the biosynthesis of the hormone ethylene, which is enhanced in response to stress conditions, such as salinity (Mizrahi, 1982).

4. Experimental

4.1. Plant growth

The tomatoes were grown in an open hydroponic drip irrigation system fully described in (Johnson et al., 2000; Johnson, 2001). Two independent systems were established enabling two treatments, a control where the irrigation water contained nutrients only and a salt treatment containing nutrients as in the control plus supplementary NaCl at a concentration of 0.4% w/v.

The plants were grown in troughs containing perlite and were irrigated with Solufeed F (Solufeed Ltd., The Riddings, Burgh Hill, Etchingham, Sussex, UK) and supplementary calcium nitrate (Ca_2NO_3) (BDH). Typically the final Ca_2NO_3 concentration was 3.8 mM although the levels of nutrients applied depended on the growth stage of the plants, which were altered in accordance with the Solufeed users manual. A digital timer controlled the irrigation of the plants and sufficient solution was applied to ensure run-through after each 'fertigation' period.

Seeds were germinated in Levingtons's Universal compost (Radnor Garden Supplies, Llandrindod, Wales) in a covered tray to maintain darkness at 25 °C. When hypocotyls emerged the cover was removed. The seedlings were transplanted when approximately 7 day old into John Innes No 2 compost and then were transplanted into the hydroponic system when a majority had developed the 5th true leaf. The chronological age varied but physiological growth stage was comparable between the two varieties. The compost root-ball was placed in a mesh pot, which had been sunk into the trough (both the pot and trough contained perlite). Over time the roots grew out of the mesh pot and into the trough. Drip irrigators were placed in the pot and in the trough. The troughs were covered with black/white co-extruded plastic to maintain the roots in darkness and to minimise algal growth.

4.2. Growth analysis

Plants were harvested from the hydroponic system at 7 and 14 day after the start of the salt treatment. Five plants were randomly selected from each treatment. The following growth parameters were recorded: leaf, stem and root fresh and dry weights, leaf surface area, stem height, and number of true leaves. Dry weights were obtained by placing the samples in an oven at 75 °C (±3 °C) for a minimum of 48 h until a constant dry weight was obtained (to four decimal places). These data were then used to calculate the relative growth rate (RGR) and the net assimilation rate (NAR) (Hunt, 1982, 1990).

4.3. Fruit tissue preparation

Fruits were harvested from the 1st and 2nd trusses when at the fully ripe stage, defined as stage 10 on the OECD tomato ripening chart (Organisation for Economic Co-operation and Development, Scheme for the application of internal standards for fruit and vegetables: www.oecd.org/agr/code/cont-e.htm). Fruits were selected for uniformity to maximise homogeneity between samples. The fruits from Edkawy and Simge F1 were analysed in two separate experiments. In the experiment investigating Edkawy, 10 fruits were selected from the control plants and 10 from the salt-treated plants. In the experiment on Simge F1, six fruits were harvested from each treatment. All the fruits were prepared for analysis using the same methodology. The fruits were cut into segments and the endocarp and the seeds were removed. The exocarp and mesocarp (flesh) was crushed using a garlic press and the remaining skin was discarded. The pulp was homogenised using a Polytron blender at speed 5 for 1 min. The samples were snap-frozen in liquid nitrogen in 1 ml aliquots and then stored at −70 °C until required.

4.4. Fourier transform infrared spectroscopy

The samples were thawed at room temperature and mixed prior to loading into the drilled wells on an aluminium plate for analysis by FT-IR; 5 μl of sample were loaded per well. In both experiments for Edkawy and Simge F1, ten machine replicates were loaded for each tomato fruit sample. The plate was then dried at 50 °C for 45 min prior to loading onto the motorised stage of a reflectance thin-layer chromatography (TLC) accessory attached to a Bruker IFS28 FT-IR spectrometer (Bruker Ltd.) equipped with a mercury–cadmium–telluride (MCT) detector cooled using liquid N_2 as detailed in (Goodacre et al., 1998; Timmins et al., 1998).

The diffuse reflectance absorbance FT-IR spectra were collected over a wavenumber range from 4000 to 600 cm^{-1} under the control of an IBM-compatible personal computer using OPUS 2.1 software running under the IBM OS/2 Warp operating system at a resolution of approximately 3.85 cm^{-1}. The resultant spectrum for each sample contained 882 variables. Spectra were acquired at a rate of 20 s^{-1}. To improve signal-to-noise ratio, 256 spectra were co-added, that is added sequentially as each was collected.

4.5. Data analysis

Plant growth data were analysed using Analysis of Variance (ANOVA) in Excel (Microsoft Office 97). The spectral data were analyzed using MATLAB version 6.1.

4.6. Cluster analysis

Principal component analysis (PCA) is an unsupervised clustering method requiring no knowledge of the data set structure and acts to reduce the dimensionality of multivariate data whilst preserving most of the variance within it (Goodacre et al., 2000). Hence it is termed a data compression method (Martens and Naes, 1989) and was applied before discriminant function analysis (DFA) (Manly, 1994). DFA is a supervised clustering method that requires a priori knowledge of the replicate structure within the data set and seeks to minimise the within-group variance and to maximise the between-group variance (Brereton, 1992; Goodacre et al., 1998). The number of principal components (PCs) used by the DFA was optimized by cross validation, which involves forming the model on a training data set and then projecting a previously unseen set of data, the test set, onto the model (as detailed in Radovic et al., 2001). This is a cyclical process where the numbers of PCs are gradually reduced to find the optimum model. Each FT-IR dataset was divided into two groups, the training set and the test set in a ratio of 3:1.

4.7. Genetic algorithms

A genetic algorithm (GA) is an optimization method based on the principles of Darwinian selection (Goldberg, 1989; Davies et al., 2000) where, over a series of generations, a population of parameter sets evolve until an optimal, or near-optimal, solution to a given problem is found.

A population of n objects or chromosomes is created, each containing a string of numbers or binary digits representing the parameters of the problem to be optimized. The population is randomised so that n sets of 'unique' parameter values can be evaluated and assigned a *fitness* value (usually a single numerical value). Once all n fitness values have been assigned, a new generation of chromosomes are created. In order for this new generation to be *fitter* than the last, principles analogous to sexual and asexual reproduction within the population are applied (Broadhurst et al., 1997). The probability of a particular parent chromosome being selected for 'sexual reproduction' is proportional to its fitness, so chromosomes with a high fitness value will have a greater chance of selection. The process of selection followed by reproduction followed by mutation is repeated until n new chromosomes are created (i.e. a new population to replace the old). The fitness value is then evaluated for each of the new chromosomes and the whole process repeats itself. The algorithm continues until a predefined stopping criterion is reached, such as a given 'optimal' fitness value is met, a certain number of generations has passed or the chromosomes have converged to similar parameter values.

Prior to analysis by the GA the raw FT-IR data were pre-processed. Spectrum characterization was carried out in order to reduce the dimensionality of the GA search space down to the salient features of the sample spectra. The salient features (i.e. subset of important variables) were found by firstly looking for the points of local maxima, minima and points of inflexion along a spectrum representative of the total data set (in this case the mean spectrum was used), and then secondly, points of maximum gradient either side of these "stationary points" were found in order to fully define the shape of the representative spectrum. Both steps of this algorithm were carried out using a program written in-house using the *Mathworks*®MATLAB scripting language.

The pre-processing step reduced the dimensionality of the data from 882 variables per sample to 64 and 71 variables in Edkawy and Simge F1, respectively. The Genetic algorithm discriminant multiple linear regression (GA-D-MLR) wavelength selection methodology uses a GA to determine the subset of v wavenumbers, taken from the data matrix, which when applied to a discriminant multiple linear regression (D-MLR) model will optimally discriminate between control and salt-treated tomato fruit of two varieties, with the control fruits being classified as 0 and the salt-treated fruits as 1. All calculations were performed using in-house software written in C++ running under Microsoft Windows NT on an IBM-compatible PC, and full details of GA-MLR are given in (Broadhurst et al., 1997). Briefly, optimization is achieved by monitoring the residual mean square error of prediction (RMSEP) for each model.

The GA uses two-point crossover with mutation (Goldberg, 1989), operating on a population of binary-encoded chromosomes, each chromosome representing v candidate wavelengths (Broadhurst et al., 1997); in this study $v=5$. The selection of parent chromosomes for the next generation is carried out using a rank-based scheme (Whitley, 1989). No two identical candidates were allowed in a given population. A total of 50 independent GA runs were performed for each tomato variety.

Acknowledgements

We thank the EU for funding the project and all the partners in the INCO-DC programme for their help and support. RG thanks the EBS Committee of the UK BBSRC for financial support. We would like to thank Professor Yuksel Tusel, Ege University, Izmir, Turkey for the kind gift of Simge F1 seed and Professor Ayman Abou-Hadid, Ain Shams University, Cairo, Egypt for supplying the Edkawy seed. We also thank all the greenhouse staff at UWA.

References

Brereton, R.G., 1992. Multivariate Pattern Recognition in Chemometrics. Elsevier, Amsterdam.

Broadhurst, D., Goodacre, R., Jones, A., Rowland, J.J., Kell, D.B., 1997. Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry. Analytica Chimica Acta 348, 71–86.

Brown, M.M., Ho, L.C., 1993. Factors affecting calcium transport and basipetal IAA movement in tomato fruit in relation to blossom-end rot. Journal of Experimental Botany 44, 1111–1117.

Cuartero, J., Fernandez-Munoz, R., 1999. Tomato and salinity. Scientia Horticulturae 78, 83–125.

Cuartero, J., Yeo, A.R., Flowers, T.J., 1992. Selection of donors for salt-tolerance in tomato using physiological traits. New Phytologist 121, 63–69.

Davies, Z.S., Gilbert, R.J., Merry, R.J., Kell, D.B., Theodorou, M.K., Griffith, G.W., 2000. Efficient improvement of silage additives by using genetic algorithms. Applied and Environmental Microbiology 66, 1435–1443.

Degen, I.A., 1997. Tables of Characteristic Group Frequencies for the Interpretation of Infrared and Raman Spectra. Acolyte Publications, Harrow, UK.

Ellis, D.I., Broadhurst, D., Kell, D.B., Rowland, J.J., Goodacre, R., 2002. Rapid and quantitative detection of microbial spoilage of meat by Fourier transform infrared spectroscopy and machine learning. Applied and Environmental Microbiology 68, 2822–2828.

Ellis, S., Mellor, A., 1995. Soils and Environment. Routledge, London, UK.

Epstein, E., Norlyn, J.D., Rish, D.W., Kingsbury, R.W., Kelley, D.B., Cunningham, G.A., Wrona, A.F., 1980. Saline culture of crops: a genetic approach. Science 210, 399–404.

FAO, 2002a. Securing enough food from limited water supplies. Available from <http://www.fao.org/english/newsroom/news/2002/6880-en.html>.

FAO, 2002b. The salt of the earth: hazardous for food production. Available from <http://www.fao.org/english/newsroom/focus/focus1.htm>.

Fiehn, O., 2001. Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. Comparative and Functional Genomics 2, 155–168.

Flowers, T.J., Yeo, A.R., 1995. Breeding for salinity resistance in crop plants: where next? Australian Journal of Plant Physiology 22, 875–884.

Franco, J.A., Banon, S., Madrid, R., 1994. Effects of protein hydrolysate applied by fertigation on the effectiveness of calcium as a corrector of blossom-end ort in tomato cultivated under saline conditions. Scientia Horticulturae 57, 283–292.

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading, Mass.

Goodacre, R., Shann, B., Gilbert, R.J., Timmins, E.M., McGovern, A.C., Alsberg, B.K., Kell, D.B., Logan, N.A., 2000. Detection of the dipicolinic acid biomarker in bacillus spores using Curie-point pyrolysis mass spectrometry and Fourier transform infrared spectroscopy. Analytical Chemistry 72, 119–127.

Goodacre, R., Timmins, E.M., Burton, R., Kaderbhai, N., Woodward, A.M., Kell, D.B., Rooney, P.J., 1998. Rapid identification of urinary tract infection bacteria using hyperspectral whole-organism fingerprinting and artificial neural networks. Microbiology 144, 1157–1170.

Griffiths, P.R., de Haseth, J.A., 1986. Fourier Transform Infrared Spectrometry. John Wiley & Sons, New York.

Hilhorst, H.W.M., Groot, S.P.C., Bino, R.J., 1998. The tomato seed as a model system to study seed development and germination. Acta Botanica Neerlandica 47, 169–183.

Ho, L.C., Belda, R., Brown, M., Andrews, J., Adams, P., 1993.

Uptake and transport of calcium and the possible causes of blossom-end rot in tomato. Journal of Experimental Botany 44, 509–518.

Hunt, R., 1982. Plant Growth Curves: The Functional Approach to Plant Growth Analysis. Edward Arnold Ltd, London.

Hunt, R., 1990. Basic Growth Analysis. Unwin Hyman, London.

Johnson, H. E., 2001. The Effect of Salinity on Tomato Growth and Fruit Quality. PhD thesis, University of Wales, Aberystwyth, UK.

Johnson, H.E., Gilbert, R.J., Winson, M.K., Goodacre, R., Smith, A.R., Rowland, J.J., Hall, M.A., Kell, D.B., 2000. Explanatory analysis of the metabolome using genetic programming of simple, interpretable rules. Genetic Programming and Evolvable Machines 1, 243–258.

Mahmoud, M.H., El-Beltagy, A.S., Helal, R.M., Maksoud, M.A., 1986a. Tomato variety evaluation and selection for salt tolerance. Acta Horticulturae 190, 559–566.

Mahmoud, M.H., Jones, R.A., El-Beltagy, A.S., 1986b. Comparative responses to high salinity between salt-sensitive and salt-tolerant genotypes of the tomato. Acta Horticulturae 190, 533–543.

Manly, B.F.J., 1994. Multivariate Statistical Methods: A Primer. Chapman & Hall, London.

Martens, H., Naes, T., 1989. Multivariate Calibration. John Wiley & Sons, New York.

Mizrahi, Y., 1982. Effect of salinity on tomato fruit ripening. Plant Physiology 69, 966–970.

Radovic, B.S., Goodacre, R., Anklam, E., 2001. Contribution of pyrolysis mass spectrometry (Py-MS) to authenticity testing of honey. Journal of Analytical and Applied Pyrolysis 60, 79–87.

Timmins, E.M., Howell, S.A., Alsberg, B.K., Noble, W.C., Goodacre, R., 1998. Rapid differentiation of closely related *Candida* species and strains by pyrolysis-mass spectrometry and Fourier transform-infrared spectroscopy. Journal of Clinical Microbiology 36, 367–374.

Toenniessen, G.H., 1984. Review of the world food situation and the role of salt-tolerant plants. In: Staples, R.C., Toenniessen, G.H. (Eds.), Salinity Tolerance in Plants—Strategies for Crop Improvement. John Wiley & Sons, New York, pp. 399–412.

Whitley, D., 1989. The GENITOR algorithm and selection pressure: Why rank-based allocation of reproductive test is best. In: Proceedings of the Third International Conference on Genetic Algorithms. Morgan Kaufmann, San Mateo, CA, pp. 434–439.

Yoshida, H., Leardi, R., Funatsu, K., Varmuza, K., 2001. Feature selection by genetic algorithms for mass spectral classifiers. Analytica Chimica Acta 446, 483–494.