Metabolomics Vol. 1, No. 3, July 2005(© 2005)
DOI: 10.1007/s11306-005-0003-1

Novel biomarkers for pre-eclampsia detected using metabolomics and machine learning

Louise C. Kenny[a,*], Warwick B. Dunn[b], David I. Ellis[b], Jenny Myers[a], Philip N. Baker[a] and the GOPEC Consortium, and Douglas B. Kell[b,*]

[a]*The Maternal and Fetal Health Research Centre, University of Manchester, St Mary's Hospital, Whitworth Park, Manchester, M13 0JH, UK*
[b]*School of Chemistry, University of Manchester, PO Box 88, Faraday Building, Sackville Street, Manchester, M60 1QD, UK*

Received 20 May 2005; accepted 25 May 2005

Pre-eclampsia is a multi-system disorder of pregnancy with major maternal and perinatal implications. Emerging therapeutic strategies are most likely to be maximally effective if commenced weeks or even months prior to the clinical presentation of the disease. Although widespread plasma alterations precede the clinical onset of pre-eclampsia, no single plasma constituent has emerged as a sensitive or specific predictor of risk. Consequently, currently available methods of identifying the condition prior to clinical presentation are of limited clinical use. We have exploited genetic programming, a powerful data mining method, to identify patterns of metabolites that distinguish plasma from patients with pre-eclampsia from that taken from healthy, matched controls. High-resolution gas chromatography time-of-flight mass spectrometry (GC-tof-MS) was performed on 87 plasma samples from women with pre-eclampsia and 87 matched controls. Normalised peak intensity data were fed into the Genetic Programming (GP) system which was set up to produce a model that gave an output of 1 for patients and 0 for controls. The model was trained on 50% of the data generated and tested on a separate hold-out set of 50%. The model generated by GP from the GC-tof-MS data identified a metabolomic pattern that could be used to produce two simple rules that together discriminate pre-eclampsia from normal pregnant controls using just 3 of the metabolite peak variables, with a sensitivity of 100% and a specificity of 98%. Thus, pre-eclampsia can be diagnosed at the level of small-molecule metabolism in blood plasma. These findings justify a prospective assessment of metabolomic technology as a screening tool for pre-eclampsia, while identification of the metabolites involved may lead to an improved understanding of the aetiological basis of pre-eclampsia and thus the development of targeted therapies.

KEY WORDS: pre-eclampsia; mass spectrometry; GC-MS; metabolomics; machine; learning; genetic programming; prognosis; diagnosis; classification.

1. Introduction

Pre-eclampsia is an important cause of maternal morbidity and mortality. The World Health Organization estimates that worldwide over 100,000 women die from pre-eclampsia each year, and the condition has been the most important cause of maternal death in the UK over recent decades (Hibbard and Milner, 1994; Lewis, 2001). Recent CESDI reports cite 1 in 6 stillbirths and 1 in 6 sudden infant deaths as occurring in pregnancies complicated by maternal hypertension, and the condition is responsible for the occupancy of approximately 20% of special care baby unit cots (CESDI, 1998).

Although the precise aetiology of pre-eclampsia is poorly defined, there is accumulating evidence for a pathogenic model of pre-eclampsia, whereby inappropriate adaptation of the interface between the maternal vasculature and the developing placenta early in pregnancy leads to the development of a poorly perfused feto-placental unit (Pijnenborg *et al.*, 1991; Hayman *et al.*, 1999). In this model, continuing poor perfusion of the placenta is proposed to result in the secretion of a factor(s) into the maternal circulation. These factors are thought to cause "activation" of the vascular endothelium and the clinical syndrome of pre-eclampsia results from widespread changes in endothelial cell function in both small and large vessels (Rodgers *et al.*, 1988; Roberts *et al.*, 1989; Kenny *et al.*, 2002). Equivalently, or in addition, one might imagine that plasma normally contains a factor(s) that maintains standard endothelial function but which is absent in pre-eclampsia.

Thus, it is clear that as pre-eclampsia originates early in pregnancy, potential therapies are most likely to be maximally effective if commenced weeks or even months prior to the clinical presentation of the disease. There are currently several candidate pharmacological therapies under investigation. However, targeted intervention

* To whom correspondence should be addressed.
E-mail: louise.kenny@manchester.ac.uk; dbk@manchester.ac.uk

is impractical as currently available tests (such as Doppler ultrasound waveform analysis of the uterine arteries) that seek to identify the condition prior to clinical presentation have low sensitivities, and are thus of limited clinical use.

Biomarkers, including surrogate markers, are well-recognised to be of great value in human disease diagnosis (Lesko and Atkinson, 2001; Frank and Hargreaves, 2003), and functional studies at the level of gene expression (transcriptomics) and protein translation (proteomics) have recently enjoyed some success in the early detection and diagnosis of cancer and its subtypes (Golub *et al*., 1999; Petricoin *et al*., 2002).

Candidate proteins have been investigated as risk determinants for pre-eclampsia, both in isolation and in combination with other markers, but have limited sensitivity and specificity (Hayman *et al*., 1999). Pre-eclampsia is undoubtedly a multisystem disorder, and the manifestations of the disease seem unlikely to be related to a single protein. Consequently, analytical methods devised to detect specific changes miss a wide range of other substances which will be numerous and may be significantly more important as surrogate (or even aetiological) markers. Although the metabolome is certainly "complementary" to transcriptomics and proteomics, it might be seen to have special advantages. In particular, it is known from both the theory underlying metabolic control analysis (Kell and Mendes, 2000) and from experiment (Raamsdonk *et al*., 2001) that, although changes in the quantities of individual enzymes might be expected to have little effect on metabolic fluxes, they can and do have significant effects on the concentrations of numerous individual metabolites. In addition, the metabolome is further down the line from gene to function and so reflects more closely the activities of the cell or organism at a functional level. Thus, as the "downstream" result of gene expression, changes in the metabolome are expected to be amplified relative to changes in the transcriptome and the proteome (Urbanczyk-Wochniak *et al*., 2003). In addition, metabolic fluxes are not regulated by gene expression alone, and metabolites are increasingly recognised as important signalling molecules (Shi *et al*., 2003). Given recent successes in disease diagnosis using NMR analyses of the metabolome (e.g., Brindle *et al*., 2002), it was therefore of interest to enquire as to whether a metabolomics approach (Oliver *et al*., 1998; Harrigan and Goodacre, 2003; Bino *et al*., 2004; Goodacre *et al*., 2004; Kell, 2004, 2005; van der Greef *et al*., 2004; Whitfield *et al*., 2004; Brown *et al*., 2005; Kell *et al*., 2005), in which as many metabolites as possible are measured, might permit a distinction between plasma from women with pre-eclampsia and that from normal pregnant women.

Gas Chromatography-Mass Spectrometry (GC-MS) provides the high resolution separation of metabolites by gas chromatography and sensitive detection by mass spectrometry that is most appropriate for complex biological fluids (e.g., Jellum *et al*., 1981; Goodacre *et al*., 2004; Dunn and Ellis, 2005; Dunn *et al*., 2005). The variant used here employs a GC-tof-MS instrument (Fiehn *et al*., 2000) optimised using a closed-loop algorithm (O'Hagan *et al*., 2005), allowing the detection, in a non-biased manner, of up to 900 metabolite peaks in some 20 min. This allows metabolites of interest (including disease biomarkers) to be detected and quantified without *a priori* knowledge of what they are and then to determine which, if any, are significant for the problem of interest (see Kell and Oliver, 2004). In health-related fields GC-MS has been used for a number of years in a range of applications, including the diagnosis of inborn errors of metabolism (Rashed, 2001).

Fourier transform infrared (FT-IR) spectroscopy involves the observation of molecules that are excited by an infrared beam, resulting in an infrared absorbance spectrum which – as with most NMR approaches – represents a "fingerprint" characteristic of any chemical or biochemical substance (Ellis *et al*., 2002), and has been previously used in metabolome profiling (Harrigan and Goodacre, 2003). Its main advantages are that it is very rapid (taking seconds), reagentless and non-destructive. FT-IR has been applied to a wide-range of biological studies including clinical ones (Ellis *et al*., 2003). Results using FT-IR will be reported elsewhere.

Profiles generated from these techniques can contain hundreds or even thousands of data points, necessitating sophisticated analytical tools. We aimed to use these high-dimensional GC-tof-MS data to define an optimum discriminatory metabolomic pattern that would distinguish plasma from women with a known diagnosis of pre-eclampsia from plasma taken from matched controls.

2. Methods

2.1. Participants

Plasma samples were obtained from the GOPEC archive. The GOPEC study was a British Heart Foundation-funded multi-centre collaborative study involving ten University Departments of Obstetrics and Gynaecology in the UK. Within this study, 1000 "low-risk" Caucasian women who developed pre-eclampsia were recruited and sampled between 1999 and 2003. Specifically, women were included if they had a systolic blood pressure $\geq$ 140 mmHg and diastolic pressure $\geq$90 mmHg on two occasions after the 20th week of pregnancy and proteinuria $>$300 mg/L in a 24 h collection, or 500 mg/24 h. Women with chronic hypertension, a history of renal or cardiovascular disease, diabetes mellitus including gestational diabetes, three or more spontaneous abortions, a hydatidiform mole in the index or earlier pregnancy, or a multiple pregnancy, were excluded from the study. Eighty-seven women within the archive who had donated blood to the study

antenatally (after diagnosis and within a week prior to delivery) were identified and were matched with 87 normal pregnant controls for maternal age, parity and BMI and for gestational age at sampling. Controls were obtained from antenatal clinics in Manchester and Dundee, and their plasma samples were only retained for this study if they subsequently experienced an uncomplicated pregnancy.

2.2. Sample collection

Blood samples were taken at the time of recruitment. Samples were collected into pre-cooled glass tubes containing EDTA using the Vacutainer® system and immediately centrifuged at 1500 g for 15 min at 4°C. Plasma was then removed and stored in aliquots at −80°C until required. The collection and storage conditions were identical for samples taken from both patients and controls.

2.3. GC-tof-MS

Sample preparation for GC-MS analysis was performed as follows; 175 μl plasma was spiked with 50 μl internal standard solution (1.53 mg/ml succinic d_4 acid, 2.34 mg/ml malonic d_2 acid, 1.59 mg/ml glycine d_5, 0.76 mg/ml glucose $^{13}C_6$; Sigma-Aldrich, Gillingham, UK) and vortex-mixed for 15 s. Four hundred and fifty micro litres of acetonitrile (AR grade; Sigma-Aldrich, Gillingham, UK) were added followed by vortex mixing (15 s) and centrifugation (13,385 g, 15 min) to deproteinise the samples. The supernatant was transferred to an Eppendorf tube and freeze dried (HETO VR MAXI vacuum centrifuge attached to a HETO CT/DW 60E cooling trap; Thermo Life Sciences, Basingstoke, UK). Two-stage sample chemical derivatisation was performed on the dried sample. 80μl 20 mg/ml O-methylhydroxylamine solution was added and heated at 40°C for 90 min followed by addition of 80μl MSTFA (*N*-acetyl-*N*-[trimethylsilyl]-trifluoroacetamide) and heating at 40°C for 90 min. Twenty microlitres of a retention index solution (4 mg/ml *n*-decane, *n*-dodecane, *n*-pentadecane, *n*-nonadecane, *n*-docosane dissolved in hexane) was added and the samples were analysed using a Agilent 6890 N gas chromatograph and 7683 autosampler (Agilent Technologies, Stockport, UK) coupled to a LECO Pegasus III electron impact time-of-flight mass spectrometer (LECO Corporation, St Joseph, USA). Optimised instrumental conditions for serum have been described elsewhere (O'Hagan *et al*., 2005) and were used here for plasma. Initial data processing of raw data was undertaken using LECO ChromaTof v2.12 software to construct a data matrix (metabolite peak *vs*. sample no.) including response ratios (peak area metabolite/peak area succinic-d_4 internal standard) for each metabolite peak in each sample. Each sample was analysed only once.

2.4. Machine learning, statistical analyses and visualisation

GC-tof-MS data, ratioed as above, were exported together with their class memberships (pre-eclampsia or normal) as an Excel table into the program The-gmax, bio-edition (Predictive Solutions Ltd, Aberystwyth; http://www.predictivesolutions.co.uk/). The program, which encodes rules as trees and evolves them according to the general principles described elsewhere (Kell, 2002; Koza, 1992; Allen *et al*., 2004), was used according to the manufacturer's instructions. The functions used were a mixture of arithmetic (+,−,*,/, larger) and logical (≤, >, AND, NOT, OR, XOR) operators. 50% of the samples were used as a hold-out set. The program ranks and reports the usage of each individual variable in the 300 most successful rules; inspection of the results from three independent runs showed that three variables were the most important and were likely to be sufficient for the discrimination, and these and other candidate variables were analysed and visualised using the Spotfire program (http://www.spotfire.com) to assess their discriminating power by inspection (Kell *et al*., 2001). From these plots, even simpler rules were derived and are given in the text.

3. Results

3.1. Patients

A summary of the patient details for the two groups (diseased/control) are detailed in table 1.

The pre-eclampsia group had significantly raised mean arterial pressure ($p < 0.0001$), which was taken as the maximum recorded value in the 24 h immediately preceding delivery, and a significantly shorter gestation period than the normal pregnant group. Individual birthweight ratios were calculated for each pregnancy. These are dependent upon maternal ethnicity, height, weight, parity and gestation at delivery, in addition to fetal birthweight and sex (Gardosi, 1998), and were

Table 1
Demographic data for patients from whom plasma samples were taken

	Normal outcome $n=87$	Preeclampsia $n=87$
Age	30 (19–43)	31 (19–41)
Parity	0 (0–2)	0 (0–2)
BMI (weight/height^2)	25 (19–46)	26 (18–46)
Max (S) BP (mm Hg)	122 (96–147)	162 (138–220)*
Max (D) BP (mm Hg)	80 (60–93)	110 (90–140)*
Delivery gestation (weeks + days)	40 + 4 (34 + 3 to 42 + 0)	37 + 0* (26 + 3 to 41 + 1)
Birth weight (g)	3420 (2380–4420)	2410 (590–4300)*
IBR (centile)	34 (10–99)	8 (0–99)*

Median (range).
Pre-eclampsia vs normal outcome.
*$p < 0.0001$.

ranked to produce a centile. The pre-eclampsia group had babies with significantly lower birthweights and birthweight ratios than the normal pregnant group ($p < 0.05$).

3.2. Metabolome data

A typical GC-tof-MS trace of the plasma metabolome is shown in figure 1.

The main part of the figure shows a typical GC-tof-MS trace for the plasma from one of the diseased patients taken at random. The data were deconvolved using the Chroma-tof software and peaks were extracted into MS-Excel. Application of genetic programming using the program The-gmax suggested that three highly discriminatory variables were labeled peaks 403, 415 and 427, and it was clear from a 3D plot of these (figure 2) that essentially all the samples from women with pre-eclampsia could be discriminated from the samples from controls using just these three variables. Because of the manner in which the plasma samples were deproteinised and derivatised, we may be certain that these are metabolites of low MW. This figure (and the inset of figure 1) shows that indeed variable 427 was raised in the disease, while variable 415 especially and 403 (in this case) were lowered, relative to the controls.

A pair of rules that may be derived from inspection of figure 2 is as follows:

Rule 1: IF $403 < 0.035$ AND (IF $415 < 0.0005$ OR IF $427 > 0.0005$) THEN disease

Rule 2: IF $403 > 0.015$ AND (IF $415 < 0.01$) THEN disease

Rule 1 gives two false positives and no false negatives, while rule 2 has 100% sensitivity and specificity. This important result shows that practically complete discrimination (100% sensitivity and 98% specificity) of the plasma samples may be made using these two rules that feature just three metabolites. The identity of the metabolites was sought using mass spectral library searches but no certain identifications could be made at this stage. This parallels the situation commonly found in plants (Fiehn *et al*., 2000). However, in contrast to proteome fingerprinting results (Petricoin *et al*., 2002), we may be certain from their mass spectra that the identity of metabolites that are given a number is the same in all samples, and the significance of this is that a comparatively simple metabolome strategy allows the detection of this small number of discriminating metabolites and enforces one's concentration on them for further study. This is in contrast to the pattern-recognition approach, in which it is not necessary (nor usually possible) to identify the basis for any diagnostic discrimination. It is also worth noting that the concentrations of the discriminatory metabolites bear no simple relationship to each other (so one is unlikely to be a major breakdown product of another, whether *in vivo* or during sample preparation or storage, nor of any other normal plasma metabolite), and that for example, the appearance of 427 in plasma from women with pre-eclampsia is not obviously coupled to the loss of 415

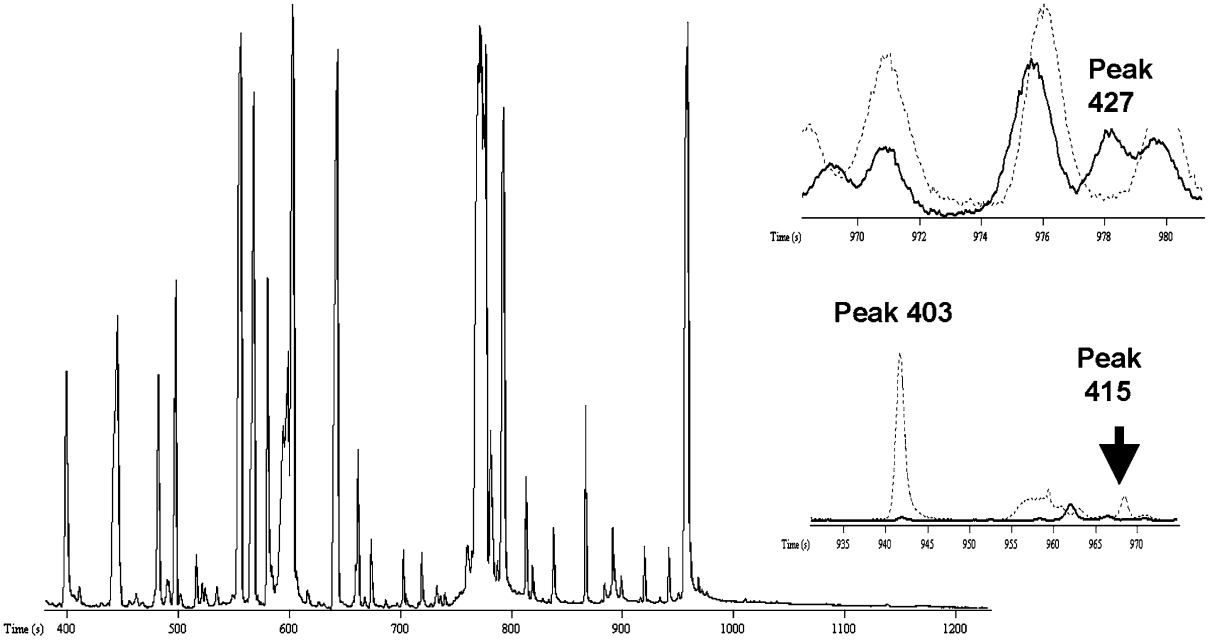

Figure 1. GC-MS total ion chromatogram for diseased patient. Inserts show the single ion monitoring chromatograms for the three peaks of interest (peak 427, peak 403 and peak 415). The full line is for a diseased patient and the dotted line for a healthy control patient. Single ion monitoring (SIM) for each metabolite was used and employed unique and highly sensitive fragment ions (peak 403, m/z 243; peak 415, m/z 202; peak 427, m/z 204 – NB these are not the molecular ion which is rarely seen with good response in this type of system). The masses described for each metabolite are the true molecular weight of the fragment ion monitored.

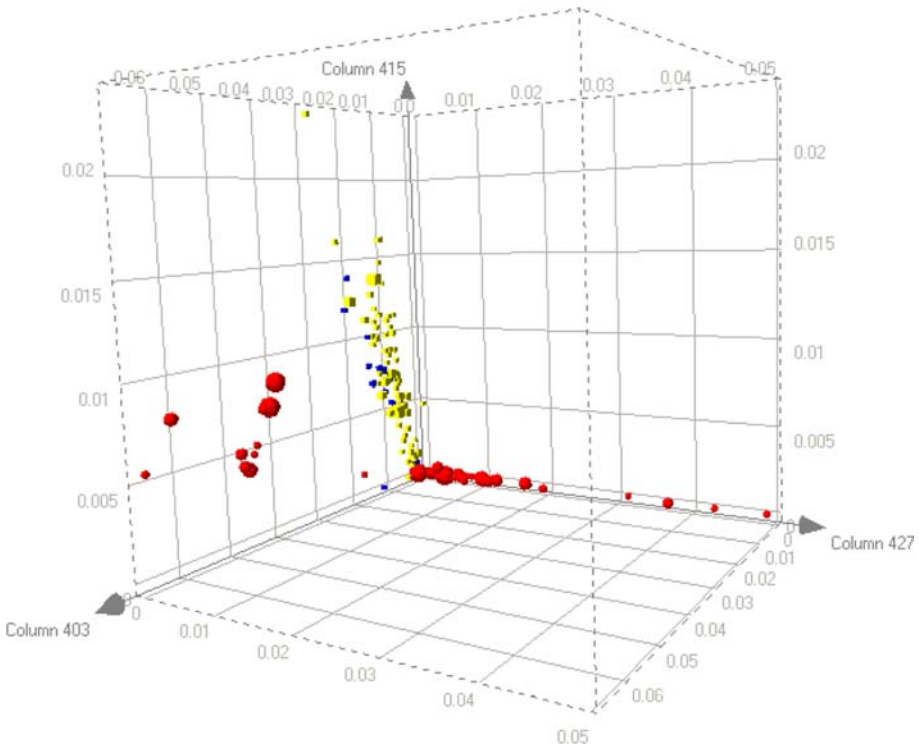

Figure 2. Discrimination of pre-eclamptic from normal plasma using just three metabolites (labelled 403, 415 and 427). Pre-eclamptic samples are encoded as red spheres, while plasma from the two control populations from Manchester (blue) and Dundee (yellow) are encoded as cubes. Another variable (metabolite peak 158) that tends to be larger in the pre-eclamptic plasma is encoded via the size of the symbols.

since 415 is normally at a noticeably lower concentration than 427. Apart from the samples where $403 > 0.03$ the concentrations of 403 and 415 are more or less collinear, suggesting a possible relationship between them. The concentration of 410 was also related to that of 415 and was only slightly less discriminatory. However, the disjoint nature of the metabolite data, requiring at least two rules to separate the classes, shows (i) that a unitary hypothesis for pre-eclampsia is inappropriate, consistent with its recognition (see above) as a multi-system disorder, and (ii) that machine learning methods (such as genetic programming) are much more suitable than is classical statistics – which tests the goodness of fit to an existing hypothesis (Breiman, 2001) – for uncovering such relationships from the data *de novo*. However, we provide a univariate statistical analysis of the distributions of these three metabolites among the patients and controls in table 2.

Since (see table 1) there are substantial differences in variables such as BP between disease and controls, one might reasonably enquire as to whether our markers are simply reflecting BP. The evidence that these metabolites are not simply BP markers comes from observing the variation between BP and the metabolites within a group, where it is clear that they are not related to each other. Figure 3 shows the data for metabolite 427.

A number of other features of the data are of interest. First, there is no obvious difference regarding the three discriminating metabolites between the two control populations from Manchester and Dundee, indicating that specific demographic factors are not responsible for these peaks. Secondly, further evidence that the loss of 415 is potentially involved in the development of disease comes from the fact (not shown) that all patients with deliveries at early gestational ages (26–28 weeks), and with higher levels of urinary protein, both indicators of disease severity, had the lowest levels of 415. None of these metabolites was significantly different in patients receiving antihypertensive therapy (and correspondingly none of them represented the metabolic products of antihypertensive drugs).

4. Discussion and conclusions

Many diseases have an uncertain aetiology, and novel strategies are required to make progress in discovering how they develop. As in functional genomics, where thousands of genes of unknown function were uncovered following the systematic genome sequencing programmes, it is now common to use data-driven expression profiling strategies in which a more specific hypothesis is the result, not the starting point, of the cycle of investigation that links hypotheses with data (Kell and Oliver, 2004).

Pre-eclampsia is such a disease, in which while there are indications that circulating factors (or maybe the lack of them) are of potential aetiological and/or

Table 2
Median and interquartile ranges (IQR) (n = 87) for the levels of three metabolites that are discriminatory between pre-eclampsia and controls

Metabolite peak	Pre-eclamptic plasma		Normal plasma		p-Value
	Median	IQR	Median	IQR	
403	0.00113	0.000535–0.00199	0.00953	0.00625–0.0128	N.S.
415	0	0–0	0.00690	0.00463–0.00989	< 0.001
427	0.00122	0–0.00623	0	0–0	< 0.001

Note: The differences for metabolites 415 and 427 were highly significant (using a Mann-Whitney U test). That for 403 was not, although by inspection and from rule 2 it clearly discriminated a subset of the samples.

diagnostic importance, we have no knowledge of what these might be. Thus according to one model (Pijnenborg *et al*., 1991; Hayman *et al*., 1999), continuing poor perfusion of the placenta is proposed to result in the secretion of one or more factors into the maternal circulation. These factors are thought to cause ''activation'' of the vascular endothelium and the clinical syndrome of pre-eclampsia results from widespread changes in endothelial cell function in both large and small vessels (Rodgers *et al*., 1988; Roberts *et al*., 1989; Kenny *et al*., 2002).

By using GC-tof-MS we were able to separate and detect several hundred metabolites from both control and diseased plasma samples, and the application of genetic programming to these data indicated that the pre-eclamptic plasma could be discriminated from the matched controls on the basis of just three metabolite peaks (two of which tended to be lower and one tended to be higher in the samples from women with pre-eclampsia, and to a certain extent this correlated with the severity of the disease). In this context it is worth commenting that the GP type of approach is to be

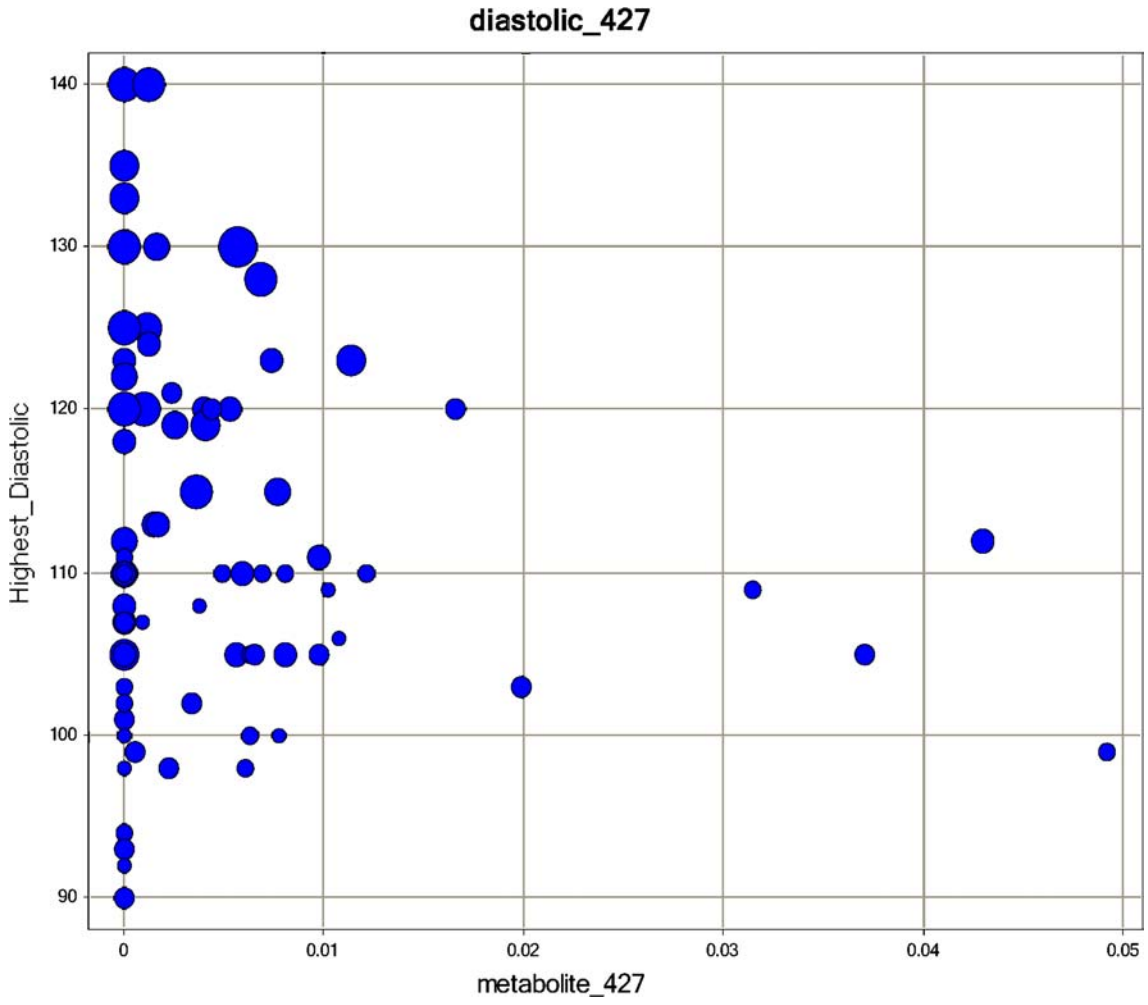

Figure 3. Lack of relationship between diastolic blood pressure and biomarker metabolite 427 in the plasma of pre-eclamptic women. In this plot the systolic blood pressure, which is fairly well correlated with the diastolic, is encoded in the size of the symbols.

preferred over other machine learning methods such as neural networks and support vector machines, as it allows one to understand the problem in terms of small subsets of input variables that it combines into rules.

In the present case, it has not yet proved possible to identify these molecules chemically, and it is clear that the next stage of a subsequent investigation is to do so. While these substances that we have identified might simply be biomarkers, it is at least possible that they are among the circulating factors that are implicated in disease aetiology. We note that the fact that 415 is lowered (effectively absent) in the disease means that attempts to purify it from pre-eclamptic plasma are likely to prove unrewarding. It could be viewed as a possible protective factor against the development of pre-eclampsia.

In the present case, only 10 each of the disease and control samples were taken at a gestational age of under 30 weeks, and a clear task for the future is to establish the extent to which these diagnostic rules apply earlier in pregnancy and thus are of greater prognostic value.

In conclusion, however, this is the first study that has identified a small subset of small-MW metabolites that effectively detects pre-eclampsia in human plasma; the potential of such metabolomic strategies in medicine is clearly considerable.

5. Conflict of interest statement.

DBK is a Director of Predictive Solutions, Ltd., the producer of the Genetic Programming software used in this study.

Acknowledgments

LK, JM, & PB thank Tommy's, the Baby Charity, for financial support. DBK thanks the BBSRC, EPSRC and the Royal Society of Chemistry for financial support. The authors wish to thank Jenny Robinson and Dympna Tansinda for their assistance with the collection of control plasma samples. We thank an anonymous referee for a useful comment.

References

Allen, J., Davey, H.M., Broadhurst, D., Rowland, J.J., Oliver, S.G. and Kell, D.B. (2004). Discrimination of the modes of action of antifungal substances by use of metabolic footprinting. *Appl. Env. Micr.* **70**, 6157–6165.

Bino, R.J., Hall, R.D., Fiehn, O., Kopka, J., Saito, K., Draper, J.Nikolau, B.J., Mendes, P., Roessner-Tunali, U., Beale, M.H., Trethewey, R.N., Lange, B.M., Wurtele, E.S. and Sumner, L.W. (2004). Potential of metabolomics as a functional genomics tool. *Trends Plant Sci* **9**, 418–425.

Breiman, L. (2001). Statistical modeling: The two cultures. *Stat Sci* **16**, 199–215.

Brindle, J.T., Antti, H., Holmes, E., Tranter, G., Nicholson, J.K., Bethell, H.W., Clarke, S., Schofield, P.M., McKilligin, E., Mosedale, D.E. and Grainger, D.J. (2002). Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics. *Nat Med* **8**, 1439–1444.

Brown, M., Dunn, W.B., Ellis, D.I., Goodacre, R., Handl, J., Knowles, J.D., O'Hagan, S., Spasic, I. and Kell, D.B. (2005). A metabolome pipeline: from concept to data to knowledge. *Metabolomics* **1**, 35–46.

CESDI (1998) Confidential Enquiry into stillbirths and deaths in infancy. 5th Annual Report. Maternal and Child Health Research Consortium.

Dunn, W.B. and Ellis, D.I. (2005). Metabolomics: current analytical platforms and methodologies. *Trends Anal Chem* **24**, 285–294.

Dunn, W.B., Bailey, N.J.C. and Johnson, H.E. (2005). Measuring the metabolome: current analytical technologies. *Analyst* **130**, 606–625.

Ellis, D.I., Broadhurst, D., Kell, D.B., Rowland, J.J. and Goodacre, R. (2002). Rapid and quantitative detection of the microbial spoilage of meat by Fourier transform infrared spectroscopy and machine learning. *Appl Environ Microbiol* **68**, 2822–2828.

Ellis, D.I., Harrigan, G.G. and Goodacre, R. (2003). Metabolic Fingerprinting with Fourier Transform Infrared Spectroscopy in Harrigan, G.G. and Goodacre, R. (Eds), *Metabolic Profiling: Its role in Biomarker Discovery and Gene Function Analysis*. Kluwer Academic, Boston.

Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, R.N. and Willmitzer, L. (2000). Metabolite profiling for plant functional genomics. *Nat Biotechnol* **18**, 1157–1161.

Frank, R. and Hargreaves, R. (2003). Clinical biomarkers in drug discovery and development. *Nat Rev Drug Discov* **2**, 566–580.

Gardosi, J. (1998). The application of individualised fetal growth curves. *J Perinat Med* **26**, 137–142.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.

Goodacre, R., Vaidyanathan, S., Dunn, W.B., Harrigan, G.G. and Kell, D.B. (2004). Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol* **22**, 245–252.

van der Greef, J., Stroobant, P. and van der Heijden, R. (2004). The role of analytical sciences in medical systems biology. *Curr Opin Chem Biol* **8**, 559–565.

Harrigan, G.G. and Goodacre, R. (Eds) (2003). *Metabolic profiling: its role in biomarker discovery and gene function analysis*. Kluwer Academic Publishers, Boston.

Hayman, R., Brockelsby, J., Kenny, L. and Baker, P. (1999). Pre-eclampsia: the endothelium, circulating factor(s) and vascular endothelial growth factor. *J Soc Gynecol Investig* **6**, 3–10.

Hibbard, B. and Milner, D. (1994). Reports on confidential enquiries into maternal deaths: an audit of previous recommendations. *Health Trends* **26**, 26–28.

Jellum, E., Bjornson, I., Nesbakken, R., Johansson, E. and Wold, S. (1981). Classification of human cancer cells by means of capillary gas chromatography and pattern recognition analysis. *J Chromatogr* **217**, 231–237.

Kell, D.B. (2002). Genotype:phenotype mapping: genes as computer programs. *Trends Genet* **18**, 555–559.

Kell, D.B. (2004). Metabolomics and systems biology: making sense of the soup. *Curr Opin Microbiol* **7**, 296–307.

Kell, D.B. (2005) Metabolomics, machine learning and modelling: towards an understanding of the language of cells. *Biochem Soc Trans* **33**, in press.

Kell, D.B. and Mendes, P. (2000). Snapshots of systems:metabolic control analysis and biotechnology in the post-genomic era in Cornish-Brown, A. and Cárdenas, M.L. (Eds), *Technological and Medical Implications of Metabolic Control Analysis*. Kluwer

Academic, Boston, pp. 3–25(and see http://dbk.ch.umist.ac.uk/WhitePapers/mcabio.htm).

Kell, D.B. and Oliver, S.G. (2004). Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays* **26**, 99–105.

Kell, D.B., Darby, R.M. and Draper, J. (2001). Genomic computing: explanatory analysis of plant expression profiling data using machine learning. *Plant Physiol* **126**, 943–951.

Kell D.B., Brown M., Davey H.M., Dunn W.B., Spasic I. and Oliver S.G. (2005) Metabolic footprinting and Systems Biology: the medium is the message. *Nat Rev Microbiol*, July issue, in press.

Kenny, L.C., Baker, P.N., Kendall, D.A., Randall, M.D. and Dunn, W.R. (2002). Differential mechanisms of endothelium-dependent vasodilator responses in human myometrial small arteries in normal pregnancy and pre-eclampsia. *Clin Sci (Lond)* **103**, 67–73.

Koza, J.R. (1992). *Genetic programming: on the programming of computers by means of natural selection*. MIT Press, Cambridge, MA.

Lesko, L.J. and Atkinson, A.J. Jr. (2001). Use of biomarkers and surrogate endpoints in drug development and regulatory decision making: criteria, validation, strategies. *Annu Rev Pharmacol Toxicol* **41**, 347–366.

Lewis, G. (2001). *Why women die. Report on Confidential Enquiries into Maternal Deaths in the United Kingdom 1997–1999*. Department of Health, Welsh Office, Scottish Office Department of Health, Department of Health and Social Services, Northern Ireland, London.

O'Hagan, S., Dunn, W.B., Brown, M., Knowles, J.D. and Kell, D.B. (2005). Closed-loop, multiobjective optimisation of analytical instrumentation: gas-chromatography-time-of-flight mass spectrometry of the metabolomes of human serum and of yeast fermentations. *Anal Chem* **77**, 290–303.

Oliver, S.G., Winson, M.K., Kell, D.B. and Baganz, F. (1998). Systematic functional analysis of the yeast genome. *Trends Biotechnol* **16**, 373–378.

Petricoin, E.F., Ardekani, A.M., Hitt, B.A., Levine, P.J., Fusaro, V.A., Steinberg, S.M., Mills, G.B., Simone, C., Fishman, D.A., Kohn, E.C. and Liotta, L.A. (2002). Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359**, 572–577.

Pijnenborg, R., Anthony, J., Davey, D.A., Rees, A., Tiltman, A., Vercruysse, L. and van Assche, A. (1991). Placental bed spiral arteries in the hypertensive disorders of pregnancy. *Br J Obstet Gynaecol* **98**, 648–655.

Raamsdonk, L.M., Teusink, B., Broadhurst, D., Zhang, N., Hayes, A., Walsh, M.C., Berden, J.A., Brindle, K.M., Kell, D.B., Rowland, J.J., Westerhoff, H.V., van Dam, K. and Oliver, S.G. (2001). A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat Biotechnol* **19**, 45–50.

Rashed, M.S. (2001). Clinical applications of tandem mass spectrometry: ten years of diagnosis and screening for inherited metabolic diseases. *J Chromatogr B Biomed Sci Appl* **758**, 27–48.

Roberts, J.M., Taylor, R.N., Musci, T.J., Rodgers, G.M., Hubel, C.A. and McLaughlin, M.K. (1989). Preeclampsia: an endothelial cell disorder. *Am J Obstet Gynecol* **161**, 1200–1204.

Rodgers, G.M., Taylor, R.N. and Roberts, J.M. (1988). Preeclampsia is associated with a serum factor cytotoxic to human endothelial cells. *Am J Obstet Gynecol* **159**, 908–914.

Shi, Y., Evans, J.E. and Rock, K.L. (2003). Molecular identification of a danger signal that alerts the immune system to dying cells. *Nature* **425**, 516–521.

Urbanczyk-Wochniak, E., Luedemann, A., Kopka, J., Selbig, J., Roessner-Tunali, U., Willmitzer, L. and Fernie, A.R. (2003). Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Rep* **4**, 989–993.

Whitfield, P.D., German, A.J. and Noble, P.J. (2004). Metabolomics: an emerging post-genomic tool for nutrition. *Br J Nutr* **92**, 549–555.