Closed-Loop, Multiobjective Optimization of Two-Dimensional Gas Chromatography/Mass Spectrometry for Serum Metabolomics

ARTICLE *in* ANALYTICAL CHEMISTRY · FEBRUARY 2007

Impact Factor: 5.64 · DOI: 10.1021/ac061443+ · Source: PubMed

Warwick B Dunn

University of Birmingham

124 PUBLICATIONS **6,583** CITATIONS

Joshua Damian Knowles

University of Birmingham

157 PUBLICATIONS **9,155** CITATIONS

David Broadhurst

Edith Cowan University

91 PUBLICATIONS **5,184** CITATIONS

Douglas Kell

The University of Manchester

586 PUBLICATIONS **26,917** CITATIONS

Closed-Loop, Multiobjective Optimization of Two-Dimensional Gas Chromatography/Mass Spectrometry for Serum Metabolomics

Steve O' Hagan, Warwick B. Dunn, Joshua D. Knowles, David Broadhurst, Rebecca Williams, Jason J. Ashworth, Maureen Cameron, and Douglas B. Kell*

School of Chemistry, The University of Manchester, Faraday Building, Sackville Street, Manchester M60 1QD, UK, and The Manchester Interdisciplinary Biocentre, The University of Manchester, 131 Princess Street, Manchester M1 7DN, UK

Metabolomics seeks to measure potentially all the metabolites in a biological sample, and consequently, we need to develop and optimize methods to increase significantly the number of metabolites we can detect. We extended the closed-loop (iterative, automated) optimization system that we had previously developed for one-dimensional GC-TOF-MS (O'Hagan, S.; Dunn, W. B.; Brown, M.; Knowles, J. D.; Kell, D. B. *Anal. Chem.* 2005, 77, 290−303) to comprehensive two-dimensional (GC×GC) chromatography. The heuristic approach used was a multiobjective version of the efficient global optimization algorithm. In just 300 automated runs, we improved the number of metabolites observable relative to those in 1D GC by some 3-fold. The optimized conditions allowed for the detection of over 4000 raw peaks, of which some 1800 were considered to be real metabolite peaks and not impurities or peaks with a signal/noise ratio of less than 5. A variety of computational methods served to explain the basis for the improvement. This closed-loop optimization strategy is a generic and powerful approach for the optimization of any analytical instrumentation.

There is increasing interest in the measurement of nominally "all" the metabolites in a sample, i.e., the metabolome.[1−5] In practice, the very wide chemical and physical nature of these metabolites[6,7] means that only a subset, a metabolic profile, is determined using a given technique.[8,9] Nevertheless, from the philosophical point of view, in which we use these methods principally for hypothesis generation rather than hypothesis testing,[10] we do seek methods that can measure as many of the metabolites as possible to maximize the biological information obtained.

Of methods currently in use,[9,11,12] those that employ a separation step coupled to mass spectrometric detection are pre-eminent. While advances in capillary electrophoresis[13,14] and LC[15,16] are making them more technically competitive, both pioneering (e.g., refs 17−20) and more recent studies (e.g., refs 21 and 22) have favored gas chromatography as being the most highly resolving technique, and thus the separation method of choice.

Despite this long history, gas chromatographic separations have been far from optimized. The reason for this is that a comparatively large set (m) of instrumental parameters may be varied, and the number of combinations varies exponentially with m such that if each can take n values the number of possible experiments is n^m. For even modest values of n and m, exhaustive search becomes impossible. Notwithstanding, it has long been known that comparatively small changes in experimental conditions can have rather substantial effects on chromatographic performance, especially in liquid chromatography.[23−25] The same is also true for electrospray ionization mass spectrometry (e.g.,

* To whom correspondence should be addressed. Phone: 0044 161 306 4492. E-mail: dbk@manchester.ac.uk; www.dbkgroup.org.

(1) Oliver, S. G.; Winson, M. K.; Kell, D. B.; Baganz, F. *Trends Biotechnol.* **1998**, *16*, 373−378.

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

(3) Tomita, M., Nishioka, T., Eds. *Metabolomics: the frontier of systems biology*; Springer: Tokyo, 2005.

(4) Vaidyanathan, S., Harrigan, G. G., Goodacre, R., Eds. *Metabolome analyses: strategies for systems biology*; Springer: New York, 2005.

(5) van der Greef, J.; Stroobant, P.; van der Heijden, R. *Curr. Opin. Chem. Biol.* **2004**, *8*, 559−565.

(6) Nobeli, I.; Ponstingl, H.; Krissinel, E. B.; Thornton, J. M. *J. Mol. Biol.* **2003**, *334*, 697−719.

(7) Nobeli, I.; Thornton, J. M. *Bioessays* **2006**, *28*, 534−545.

(8) Fiehn, O. *Comp. Funct. Genomics* **2001**, *2*, 155−168.

(9) Goodacre, R.; Vaidyanathan, S.; Dunn, W. B.; Harrigan, G. G.; Kell, D. B. *Trends Biotechnol.* **2004**, *22*, 245−252.

(10) Kell, D. B.; Oliver, S. G. *Bioessays* **2004**, *26*, 99−105.

(11) Dunn, W. B.; Ellis, D. I. *Trends Anal. Chem.* **2005**, *24*, 285−294.

(12) Dunn, W. B.; Bailey, N. J. C.; Johnson, H. E. *Analyst* **2005**, *130*, 606−625.

(13) Soga, T.; Ueno, Y.; Naraoka, H.; Ohashi, Y.; Tomita, M.; Nishioka, T. *Anal. Chem.* **2002**, *74*, 2233−2239.

(14) Soga, T.; Ohashi, Y.; Ueno, Y.; Naraoka, H.; Tomita, M.; Nishioka, T. *J. Proteome Res.* **2003**, *2*, 488−494.

(15) Plumb, R.; Castro-Perez, J.; Granger, J.; Beattie, I.; Joncour, K.; Wright, A. *Rapid Commun. Mass Spectrom.* **2004**, *18*, 2331−2337.

(16) Wilson, I. D.; Nicholson, J. K.; Castro-Perez, J.; Granger, J. H.; Johnson, K. A.; Smith, B. W.; Plumb, R. S. *J. Proteome Res.* **2005**, *4*, 591−598.

(17) Horning, E. C.; Horning, M. G. *Clin. Chem.* **1971**, *17*, 802−809.

(18) Tanaka, K.; West-Dull, A.; Hine, D. G.; Lynn, T. B.; Lowe, T. *Clin. Chem.* **1980**, *26*, 1847−1853.

(19) Jellum, E.; Bjornson, I.; Nesbakken, R.; Johansson, E.; Wold, S. *J. Chromatogr.* **1981**, *217*, 231−237.

(20) Greenaway, W.; May, J.; Scaysbrook, T.; Whatley, F. R. *Z. Naturforsch., C* **1991**, *46*, 111−121.

(21) Fiehn, O.; Kopka, J.; Dormann, P.; Altmann, T.; Trethewey, R. N.; Willmitzer, L. *Nat. Biotechnol.* **2000**, *18*, 1157−1161.

10.1021/ac061443+ CCC: $37.00
Published on Web 12/15/2006

refs 26−28). Another issue, of course, is that we do not know the numbers of metabolites that might be present in a given matrix, although for serum we would argue that values for the number of native metabolites in the decade 1−10 000 seem reasonable based on a combination of our knowledge of the major metabolic pathways (e.g., refs 29−32) and what has been observed in the more modern experiments designed to explore this question (e.g., ref 33). We note too the potential contribution of the gut microflora to the serum metabolome,[34] which may be more pronounced in urine.[16]

In recent work,[35] inspired by the "Robot Scientist" idea,[36] we exploited a closed-loop method in which we automated and iterated the entire process of parameter setting, performance of the GC run, analysis of the data obtained (in terms of peak number, run time, and a metric of signal/noise ratio), and changing of the parameter set, with the result that with just 240 runs we could improve an already excellent method 3-fold in terms of the number of peaks detected. A multiobjective evolutionary algorithm, PESA-II ,[35,37,38] was used as the heuristic for navigating the search space of some 200 000 000 combinations.

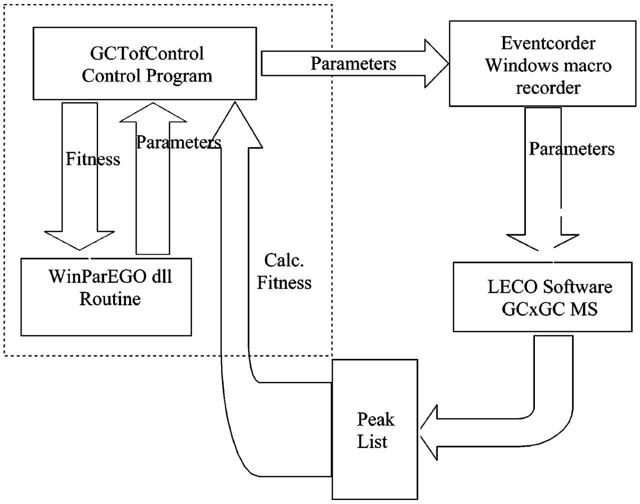


Figure 1. Closed-loop control of GC×GC-MS parameters using the ParEGO algorithm.

Comprehensive two-dimensional gas chromatography (GC×GC)[39−44] describes a general method in which substances eluting from a first column (typically nonpolar) over a certain time window are focused and then released (via a process termed modulation) on to a second (typically more polar) column where they are further separated. The technique has the potential for increasing further the number of peaks determined in metabolomics experiments caused by a combined effect of increased peak capacities, chromatographic resolution and signal-to-noise ratios.[45] One distinct advantage is that metabolites of the same volatility, and hence the same retention time on column 1, can be resolved chromatographically on column 2 if their polarities are different, something that is not achievable with 1D GC and that, consequently, decreases the reliance on deconvolution software. With two columns, there are even more experimental parameters that can be varied, and correspondingly, it has not been subjected to extensive optimization, let alone the closed-loop optimization of the type described above. It was therefore of interest to develop a comprehensive two-dimensional GC×GC method that would maximize the number of metabolites we could observe, this time using human serum. This paper describes a successful implementation of closed-loop optimization in this system, leading to a method that produces more than 4000 peaks corresponding to almost 2000 different metabolite peaks.

EXPERIMENTAL SECTION

Biological Information. Sample Preparation. Deproteinization of human serum (pooled serum from 17 individuals therefore representing the optimal metabolome expected; Sigma-Aldrich, Gillingham, UK) was performed by addition of 600 μL of methanol (AR Grade, Sigma-Aldrich, Gillingham, UK) to 200 μL of serum followed by vortex mixing (15 s), centrifugation (15 min, 13385*g*), and lyophiliation of the supernatant (HETO VR MAXI vacuum centrifuge attached to a HETO CT/DW 60E cooling trap; Thermo Life Sciences, Basingstoke, UK). A two-stage chemical derivatization procedure was performed. A 50-μL aliquot of 20 mg/mL *O*-methylhydroxylamine solution was added and heated at 40 °C for 80 min followed by addition of 50 μL of *N*-acetyl-*N*-(trimethylsilyl)trifluoroacetamide and heating at 40 °C for 80 min. All sample solutions were analyzed within 36 h of derivatization.

Optimization. The GC×GC-MS instrument (Agilent 6890N gas chromatograph (Agilent Technologies, Stockport, UK) and Gerstel MPS2L autosampler (Gerstel, Baltimore, MD) coupled

[22] Catchpole, G. S.; Beckmann, M.; Enot, D. P.; Mondhe, M.; Zywicki, B.; Taylor, J.; Hardy, N.; Smith, A.; King, R. D.; Kell, D. B.; Fiehn, O.; Draper, J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 14458−14462.

[23] Glajch, J. L.; Kirkland, J. J.; Snyder, L. R. *J. Chromatogr.* **1982**, *238*, 269−280.

[24] Kirkland, J. J.; Glajch, J. L. *J. Chromatogr.* **1983**, *255*, 27−39.

[25] Glajch, J. L.; Kirkland, J. J.; Minor, J. M. *J. Liq. Chromatogr.* **1987**, *10*, 1727−1747.

[26] Vaidyanathan, S.; Broadhurst, D. I.; Kell, D. B.; Goodacre, R. *Anal. Chem.* **2003**, *75*, 6679−6686.

[27] Vaidyanathan, S.; Kell, D. B.; Goodacre, R. *Anal. Chem.* **2004**, *76*, 5024−5032.

[28] Moberg, M.; Markides, K. E.; Bylund, D. *J. Mass Spectrom.* **2005**, *40*, 317−324.

[29] Goto, S.; Okuno, Y.; Hattori, M.; Nishioka, T.; Kanehisa, M. *Nucleic Acids Res.* **2002**, *30*, 402−404.

[30] Duarte, N. C.; Herrgard, M. J.; Palsson, B. Ø. *Genome Res.* **2004**, *14*, 1298−1309.

[31] Kanehisa, M.; Goto, S.; Kawashima, S.; Okuno, Y.; Hattori, M. *Nucleic Acids Res.* **2004**, *32*, D277−280.

[32] Kanehisa, M.; Goto, S.; Hattori, M.; Aoki-Kinoshita, K. F.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M.; Hirakawa, M. *Nucleic Acids Res.* **2006**, *34*, D354−357.

[33] Want, E. J.; O'Maille, G.; Smith, C. A.; Brandon, T. R.; Uritboonthai, W.; Qin, C.; Trauger, S. A.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 743−752.

[34] Nicholson, J. K.; Holmes, E.; Wilson, I. D. *Nat. Rev. Microbiol.* **2005**, *3*, 431−438.

[35] O'Hagan, S.; Dunn, W. B.; Brown, M.; Knowles, J. D.; Kell, D. B. *Anal. Chem.* **2005**, *77*, 290−303.

[36] King, R. D.; Whelan, K. E.; Jones, F. M.; Reiser, P. G. K.; Bryant, C. H.; Muggleton, S. H.; Kell, D. B.; Oliver, S. G. *Nature* **2004**, *427*, 247−252.

[37] Corne, D.; Knowles, J.; Oates, M. *Lecture Notes in Computer Science,* Springer: Paris, France, 2000; pp 869−878.

[38] Corne, D.; Jerram, N. R.; Knowles, J.; Oates, M., Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2001); Morgan Kaufmann: San Francisco, CA, 2001; pp 283−290.

[39] Bertsch, W. *HRC, J. High Resolut. Chromatogr.* **2000**, *23*, 167−181.

[40] Ong, R. C.; Marriott, P. J. *J. Chromatogr. Sci.* **2002**, *40*, 276−291.

[41] Blumberg, L. M. *J. Chromatogr., A* **2003**, *985*, 29−38.

[42] Marriott, P.; Shellie, R. *Trends Anal. Chem.* **2002**, *21*, 573−583.

[43] van Mispelaar, V. G.; Tas, A. C.; Smilde, A. K.; Schoenmakers, P. J.; van Asten, A. C. *J. Chromatogr., A* **2003**, *1019*, 15−29.

[44] Harynuk, J.; Marriott, P. J. *Anal. Chem.* **2006**, *78*, 2028−2034.

[45] Welthagen, W.; Shellie, R. A.; Spranger, J.; Ristow, M.; Zimmermann, R.; Fiehn, O. *Metabolomics* **2005**, *1*, 65−73.

Table 1. Instrument Parameters for Optimization Showing Final Ranges Used[a]

variable parameters	units	min	max	inc	step
sample volume	μL	1.0	5	1.0 (0.1)	5 (45)
corrected column flow	mL/min	0.8	2	0.2 (0.1)	7 (12)
split ratio	-	1:10	1:80	5 (0.1)	15 (700)
inlet temperature	°C	200	280	10 (1)	9 (40)
oven 1 start hold time	min	3	8	1 (0.01)	6 (500)
oven 1 ramp rate	°C/min	5	26	3 (0.1)	8 (95)
oven 1 final temperature	°C	260	300	10 (1)	5 (40)
oven 1 final hold time	min	0	5	1 (0.01)	6 (400)
oven 2 start temperature	°C	55	75	4 (0.01)	6 (2000)
transfer line temperature	°C	220	280	20 (1)	4 (60)
modulator temperature offset	°C	15	55	10 (1)	5 (40)
second dimension time	min	4	7	1 (0.0001)	4 (30000)
hot pulse time	sec	0.2	0.5	0.1 (0.01)	4 (30)
acquisition rate	Hz	30	160	10 (1)	14 (440)
ion source temperature	°C	220	280	20 (1)	4 (60)
fixed parameters	units	default			
oven 1 start temperature	°C	50			
initial detector voltage	V	1700			

[a] Increments and steps for experiments 1−217 are shown wihout parentheses, and for experiments 218 and higher are shown in parentheses. Oven 2 start hold time, ramp speed, final temperature, and final temperature hold time were identical to oven 1 parameters. It will be noted that the final search space, which is the product of the numbers in the right-hand column of the variable parameters, was 7.32×10^9 for the initial search (and $\sim 9.8 \times 10^{32}$ for the quasi-continuous search).

Table 2. Objectives or Fitness Functions Used.

objective	description	optimization direction
PeakCount	peak count after adjustment for noise and duplicates	maximize
RTM	run time	minimize
ANND	average nearest peak—peak neighbor distance of 25% worst peaks	maximize
ASN	average signal-to-noise of 10% worst peaks	maximize

to a Leco Pegasus IV time-of-flight (TOF) mass spectrometer (Leco Corp., St. Joseph, MO)) that we used exploit a four-jet nitrogen-based cryogenic modulation system. Columns 1 and 2 were, respectively, DB-1 (30 m × 250 μm × 0.25 μm; Agilent J&W Scientific) and BPX-50 (1.5 m × 100 μm × 0.1 μm; SGE, Milton Keynes). The Windows-based ChromaTof v2.25 software was obtained from the manufacturer and ran on an IBM-compatible PC. It was employed for instrument control and raw data processing, including chromatographic deconvolution, but does not have a suitable application programming interface by which an external control program could be used to automate acquisition and processing of GC×GC-MS data. Thus, to automate this process, we are required to "mimic" manual operator input. To achieve this, a Windows macro recorder, Eventcorder (http://www.eventcorder.com/), was used to capture and play back manual keyboard/mouse movements. As well as its own scripting environment, Eventcorder itself has its own ActiveX API and therefore playback can be controlled via any ActiveX aware programming language, such as Visual Basic, Delphi, etc. This enables playback of recorded scripts under program control with variable text (keystrokes) sent to the client program, as well as access to the full suite of features provided by the programming language.

To control Eventcorder (and hence the LECO software), we (S.O'H.) developed a Microsoft VB6 Program, GCTofControl V1.8; this software also acted as the interface to the heuristic algorithm (WinParEGO) used for choosing the parameters, read in the Leco peak list exported data files (as ASCII CSV format) and calculated fitness values from these.

The algorithm used, WinParEGO, is a C/C++ dynamic link library based on the command line ParEGO implementation developed by J.D.K. and ported from Linux to Windows. The ParEGO algorithm[46] is a multiobjective version of the efficient global optimization (EGO) algorithm of Jones and colleagues .[47] It uses a design and analysis of computer experiment (DACE)[48−50] approach to model the fitness landscape(s), based on an initial "Latin hypercube" sampling of the parameter space. Subsequently, the model is used to suggest the next experiment (set of instrumentation parameter values), such that the "expected improvement" in the fitness function is maximized. The notion of expected improvement implicitly ensures that ParEGO balances exploration of new parameter combinations with exploitation and fine-tuning of parameter values that have led to good "fitnesses" in previous experiments. The DACE model is updated after each fitness evaluation. The overall arrangement is given in Figure 1.

Instrument Parameters. The instrument parameters that were chosen for optimization are listed in Table 1; the oven 1

[46] Knowles, J. *IEEE Trans. Evol. Comput.* **2006**, *10*, 50−66.

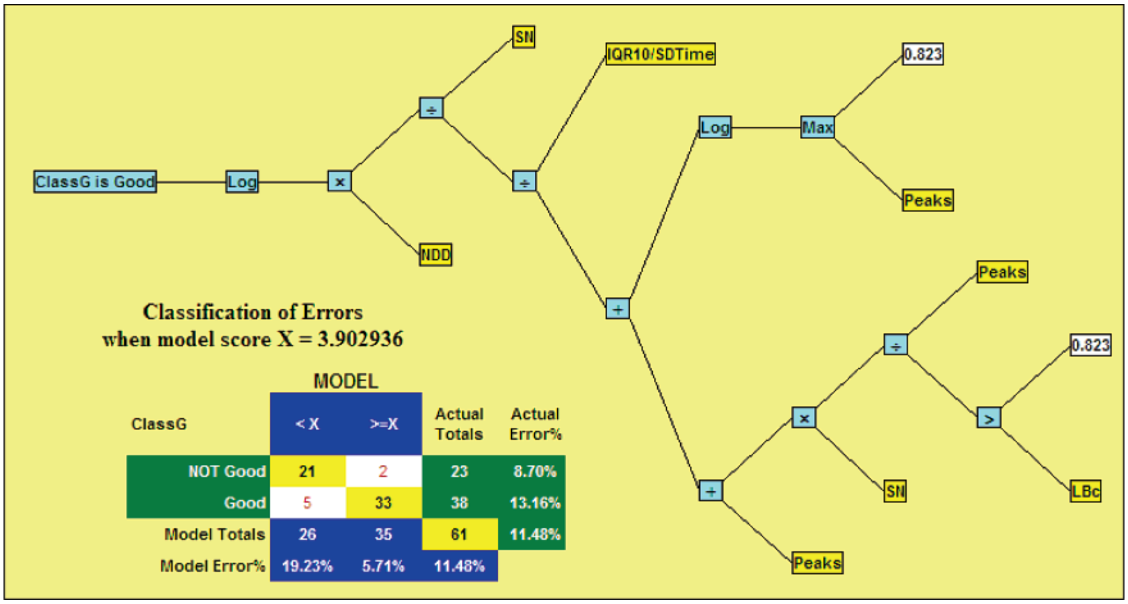
[47] Jones, D. R.; Schonlau, M.; Welch, W. J. *J. Global Opt.* **1998**, *13*, 455−492.

[48] Sacks, J.; Welch, W.; Mitchell, T.; Wynn, H. *Stat. Sci.* **1989**, *4*, 409−435.

[49] Crary, S. B. *Analog Integr. Circ. Signal Proc.* **2002**, *32*, 7−16.

[50] Chen, V. C. P.; Tsui, K. L.; Barton, R. R.; Meckesheimer, M. *IIE Trans.* **2006**, *38*, 273−291.

[51] Hicks, C. R.; Turner, K. V., Jr. *Fundamental concepts in the design of experiments,* 5th ed.; Oxford University Press: Oxford, 1999.



Classification of Errors
when model score X = 3.902936

	MODEL			
ClassG	< X	>=X	Actual Totals	Actual Error%
NOT Good	21	2	23	8.70%
Good	5	33	38	13.16%
Model Totals	26	35	61	11.48%
Model Error%	19.23%	5.71%	11.48%	

Figure 2. Auxiliary fitness function GMax. A GMax-bio model of an analyst's classification of the run quality; although not used during optimization, this provided a useful aid for subsequent analysis.

start temperature and the detector voltage were fixed for the optimization runs we carried out, but could also be set from within the GCTofControl control program. The initial oven temperature was chosen, in a preoptimization experiment to ensure that metabolites of greatest volatility were detected, since too high a temperature would mean that such metabolites will elute in the solvent front (where data are not collected to increase filament and detector lifetimes). The detector voltage was set at 1700 V, initially and an equivalent sensitivity throughout the optimization experiment was ensured by determination of the S/N for mass 69 of a calibrant gas (PFTBA) on a daily basis and adjustment of the detector voltage to maintain a similar S/N and sensitivity (4 adjustments were performed).

In the initial set of 50 experiments, parameters were allocated using the Latin hypercube approach, which attempts to distribute the parameter values across the parameter space (as in standard design of experiments strategies[51,52]). Subsequently, experimental parameters were generated using the ParEGO genetic algorithm's DACE model. The set of 50 hypercube-generated experiments gives the ParEGO algorithm sufficient data upon which to build its initial DACE model.

For experiments up to number 217, a somewhat coarser discretization of parameter ranges was used (Table 1). However, for experiments 218–246, the parameter increments were decreased to the lowest limit that would remain as acceptable inputs to the Leco software in an attempt to make the parameters approximate more nearly continuous functions. This was done because there are theoretical grounds for believing that the DACE model of the ParEGO algorithm would operate more efficiently with (more nearly) continuous values. After experiment 246, an optimal set of conditions was being approached as shown by the same parameter values being employed for consecutive experiments. The data were surveyed to describe the optimal conditions for each parameter. After optimal conditions were determined, a more targeted set of experiments were designed (by W.B.D.) to explore the local search space around these optimal conditions more systematically by experimentally varying one parameter while maintaining the other parameters at a constant level (experiments 250–300).

Fitness (Objective) Functions. In many areas of bioanalysis, it is not possible to know a priori the nature of the compounds of interest.[53] In metabolomics in particular, where it is desired to obtain information on as many (perhaps previously unknown) metabolites as possible, it is not feasible to use a targeted compound approach to analysis. Also, it is becoming important to construct experiments such that the data obtained (and the metadata[54,55]) may be reused at a future date—perhaps for an entirely unforeseen application. It is therefore prudent to capture as much analytical information as possible. For this reason, our primary "fitness function" or objective was chosen to be the number of peaks detected in the chromatogram—including appropriate filtering[56] of noise and potential duplicate peaks (see below: Data Preprocessing).

In a typical metabolomic experiment, it is likely that many hundreds if not thousands of samples will need to be analyzed, so the overall time taken to do each analysis becomes critical to the practicality of the experiment. The analytical run time was chosen as the next most important objective.

In GC×GC, the elution time chromatogram is characterized by two time dimensions, corresponding to the retention time on

[52] Montgomery, D. C. *Design and analysis of experiments,* 5th ed.; Wiley: Chichester, 2001.

[53] Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779–787.

[54] Jenkins, H.; Hardy, N.; Beckmann, M.; Draper, J.; Smith, A. R.; Taylor, J.; Fiehn, O.; Goodacre, R.; Bino, R.; Hall, R.; Kopka, J.; Lane, G. A.; Lange, B. M.; Liu, J. R.; Mendes, P.; Nikolau, B. J.; Oliver, S. G.; Paton, N. W.; Roessner-Tunali, U.; Saito, K.; Smedsgaard, J.; Sumner, L. W.; Wang, T.; Walsh, S.; Wurtele, E. S.; Kell, D. B. *Nat. Biotechnol.* **2004**, *22*, 1601–1606.

[55] Spasic, I.; Dunn, W. B.; Velarde, G.; Tseng, A.; Jenkins, H.; Hardy, N. W.; Oliver, S. G.; Kell, D. B. *BMC Bioinformatics* **2006**, *7*, 281.

[56] Brown, M.; Dunn, W. B.; Ellis, D. I.; Goodacre, R.; Handl, J.; Knowles, J. D.; O'Hagan, S.; Spasic, I.; Kell, D. B. *Metabolomics* **2005**, *1*, 35–46.

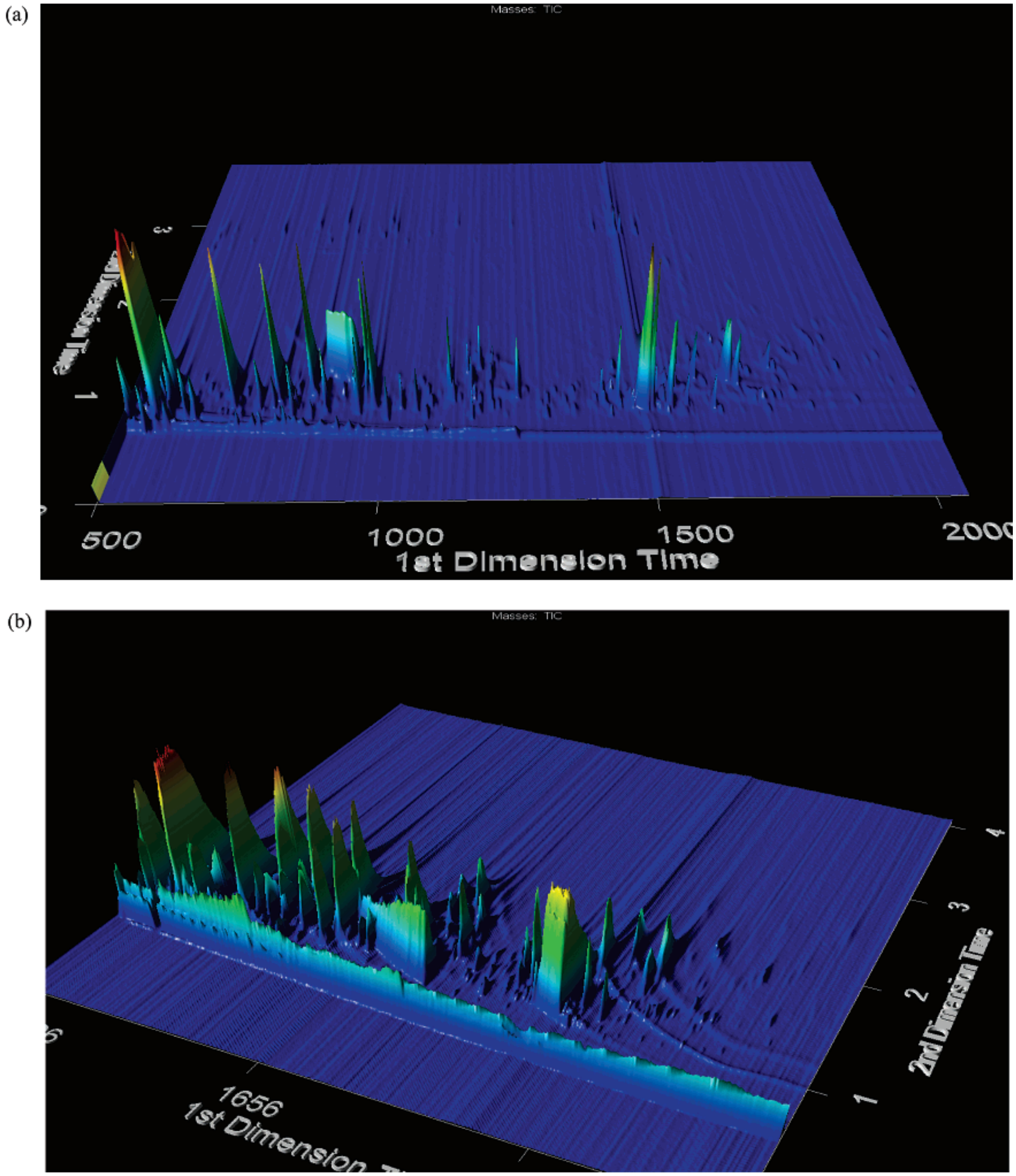


Figure 3. Typical GC×GC-TOF-MS 3D chromatograms from (a) experiment 7, one of the better runs during the first 50 experiments, and (b) from the final optimized set of conditions.

the first column and the retention time on the second column. A good interpretation of the resolution between two peaks, which takes into account both time dimensions, would be the diagonal distance between them on the 2D plane; the closest peak to any given peak will then be its nearest neighbor. For each peak, we calculated the nearest-neighbor distance after removal of noise and duplicate peaks. Ignoring the fact that each pair of peaks would be represented twice, we used the average of the lowest 25% nearest-neighbor distances as our third fitness function. In practice, because the interval between peaks on the first column is always larger than the interval between peaks on the second column, when there is more than one peak at a given first column elution time (as will happen most of the time), this fitness measure reduces to the time separation on the second column.

For the final fitness function, we chose to measure the average signal-to-noise ratio of the peaks in the run. However, as nearly all peaks had very good signal-to-noise, we limited the calculation to the average of the worst 10% of peaks, so that the fitness function would reflect improvements (if any) in the worst peaks observed—either through the reduction in noise peaks or improvement in intensity of “true” peaks. These objectives are set out in Table 2.

Data Preprocessing. The Leco software exports a text file containing a 2D peak list, that is 2D in the sense that chromatographic peaks are characterized by two retention time dimensions, that of the first GC column and that of the second GC column. Although each peak can be expanded into a full mass spectrum, and the data are therefore multidimensional, we only utilized the

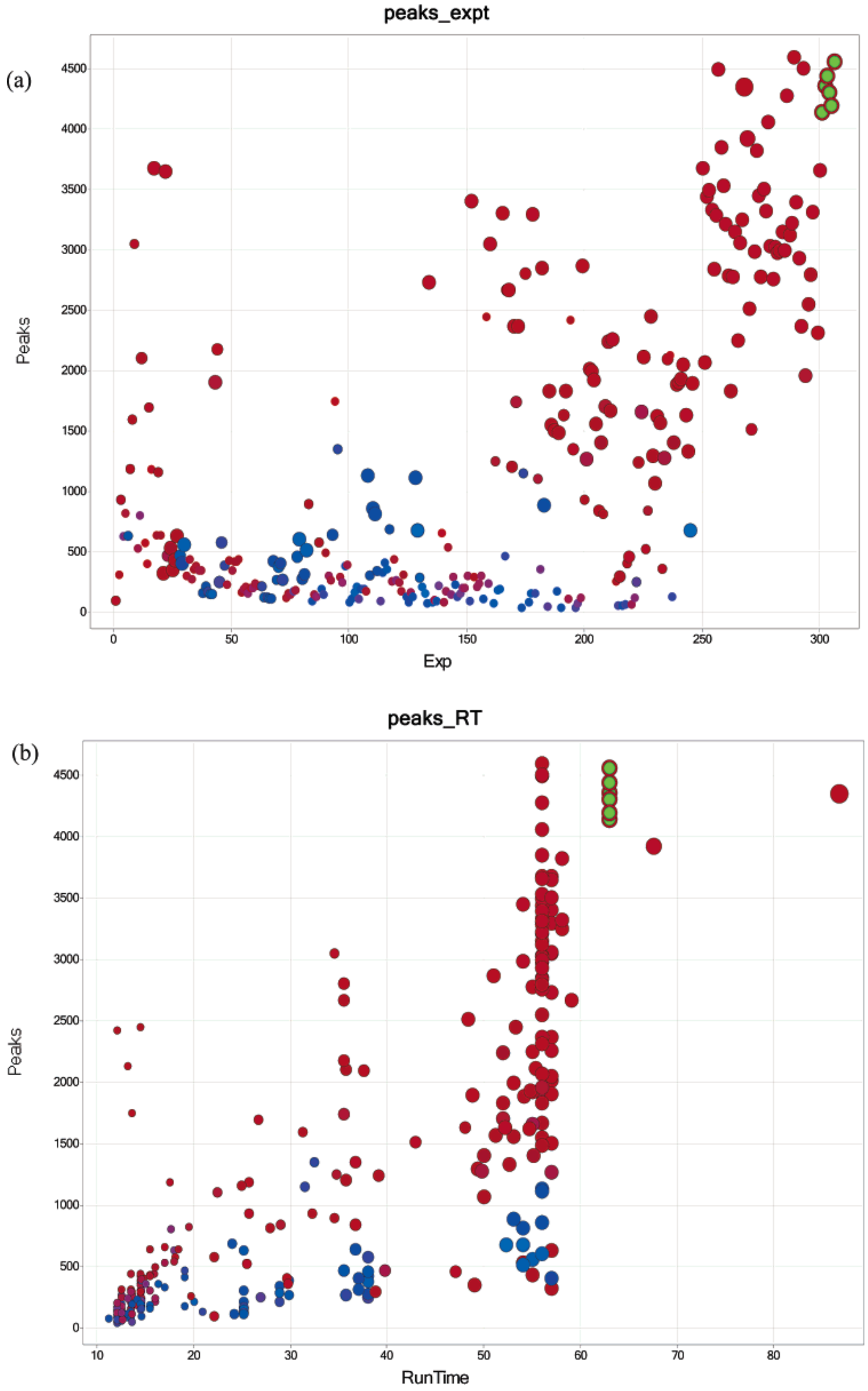


Figure 4. Improvement in peak number during the closed-loop optimization of GC×GC-TOF-MS measurements of human serum. The size of the symbol encodes the run time while the color (red low, blue high) encodes the nearest-neighbor distance defined above. The last 6 experiments in green represent validation replicates using the final chosen conditions. (a) Peaks vs experiment number. (b) Peaks vs run time.

peak area of the quantification ion identified by the peak deconvolution software for our analysis. (Note: due to the nature of the deconvolution algorithm, there is no guarantee that, for a given peak, the same quantification ion will be chosen from run to run, so this may introduce some bias; however, this is unavoidable as we have little or no control over deconvolution parameters). Due to the Leco peak detection algorithm and the way that GC "slices" from the first column are trapped and fed into the second column, there is a high potential for the occurrence of duplicate or multiplet chromatographic peaks to be identified—i.e., multiple peaks in the output file are in fact only one chemical entity. The match required to combine was set at 500, a reasonable value to ensure modulated

Table 3. Auxiliary Fitness Functions.

auxiliary fitness function	description
IQR10	interquartile range of second dimension elution time, after applying a threshold of 10 on signal-to-noise
IQR10/SDTW	IQR10 as described above divided by the width of the second dimension time window
LBc	overlap of second dimension elution time peak-count histogram with perfect rectangular distribution of same total area
LBa	overlap of second dimension elution time peak-area histogram with perfect rectangular distribution of same total area
GMax	the virtual analyst: A GMax-Bio model based on an analyst's qualitative judgment of a subset of runs

Table 4. Set of Final Optimized Conditions after 300 Experiments.

variable parameters	units	optimized value
sample volume	μL	3
corrected column flow	mL/min	1
split ratio		01:15
inlet temperature	°C	260
oven 1 start hold time	min	5
oven 1 ramp rate	°C/min	5
oven 1 final temperature	°C	290
final temperature duration	min	10
oven 2 start temperature	°C	57
oven 2 start hold time	min	5
oven 2 ramp rate	°C/min	5
oven 2 final temperature	°C	290
final temperature duration	min	10
transfer line temperature	°C	220
modulator temperature offset	°C	25
second dimension time	s	6
hot pulse time	s	0.3
acquisition rate	Hz	130
ion source temperature	°C	240

peaks of the same metabolite are combined. Another problem is that the position of the initial solvent front changes with instrument parameters (which the optimization algorithm alters from run to run), so that using a fixed solvent delay runs the risk of either losing genuine peaks if the solvent delay is too long or passing unwanted solvent peaks if the solvent delay is too short in comparison with the position of the solvent front. Both of these effects would lead to an incorrect estimate in the peak number fitness, and therefore, simple preprocessing steps were applied prior to calculating fitness.

Initial noise removal involves removing all peaks from the peak list with a peak area of <700. This equates to an approximate S/N < 5, dependent on the quantification ion employed.

To overcome the occurrence of duplicate peaks, we utilized the Leco software's mass spectral library search facility (employing NIST/EPA/NIH 02 mass spectral library and the publicly available MPI-Golm library; http://csbdb.mpimp-golm.mpg.de/csbdb/dload/dl_msri.html) to label each peak first with a tentative ID based on mass spectrum similarity measures. Thus, the primary list of duplicates constituted peaks that had been labeled with the same name more than once. However, we then applied the additional criterion that the peak separation on the first and second dimensions be within (1.5 × the second dimension time window) and 0.25 s, respectively. For the list of peaks meeting these criteria, the most intense peak was retained and the remainder were rejected as "duplicates". (Note: the fact that these two libraries provide different text identifications for the same chemical entity results in compounds identified by the two different libraries at different peak positions not being detected as a duplicate peak.) Also the match with the highest similarity was used for filtering, which in our experience is not necessarily the correct identification and a metabolite with a lower similarity match may be the correct metabolite identification. This may provide some bias to the results.

To set an appropriate solvent delay, prior to the main run, we carried out 50 experiments with varying parameters; an analyst then determined the position of the solvent front in each run. The genetic programming application GMax-Bio V2.8 (http://www.thegmax.com/) was then used to model the position of the solvent front as a function of the parameters. A reasonable model of the solvent front position in terms of the initial hold time, ramp, and gas flow rate was found. This model was incorporated into GCTofControl V1.8 as a means of setting an appropriate solvent delay. No large solvent peak, which is detrimental to filament and detector lifetime, was observed in any experiment.

Auxiliary Fitness Functions. In addition to the fitness function used during optimization, we calculated several auxiliary fitness functions that were not fed into ParEGO for optimization. These were implemented to assess potential fitness functions for future use and as a means of providing additional "figures of merit" that could be used during the analysis of the data.

All of the auxiliary fitness functions used just the peak data for the elution time on the second dimension. All except "GMax" were designed to measure the evenness of the distribution of peaks over the second dimension.

During the initial runs, it became apparent that some runs were quite poor in terms of their appearance and structure, and it was discernible by both novice and experienced analysts alike that, in fact, several of these runs appeared to have quite "abnormal" appearance. Thus, it would have been desirable to reject such runs, albeit that the fitness functions chosen did not appear to be able to detect such runs.

To try to overcome this, we used GMax-bio to model a subset of 123 runs, which had been categorized as "good", "bad", or "anomalous", although with the version of GMax-Bio we had available at the time, we were only able to model the good/not-good classification. The input parameters to the model were the instrument parameters and the other fitness functions (including the auxiliary fitness functions other than GMax). A reasonable

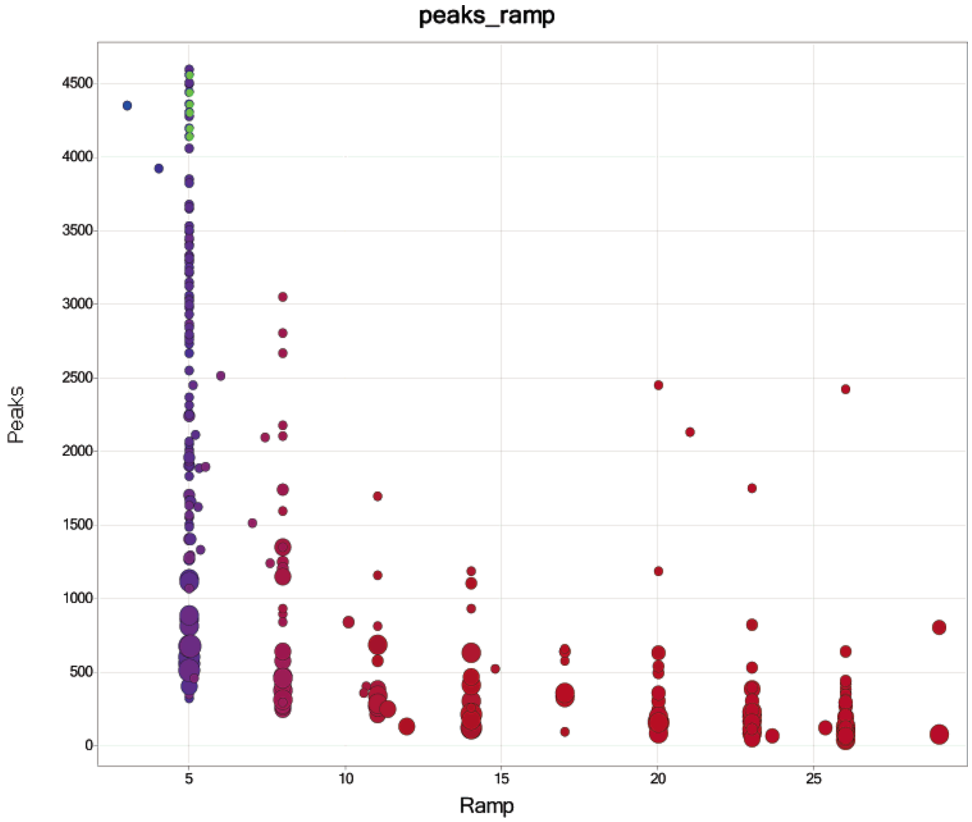


Figure 5. Effect of ramp speed (°C·min^{-1}) on the number of peaks observed during GC×GC-TOF-MS optimization. Other conditions as in Figure 4.

model was found to require only ASN, ANND, PeakCount, LBc, and IQR10 and, thus, was a nonlinear transformation of some of the other fitness functions and auxiliary fitness functions.

Although we did not carry out any further optimizations using these auxiliary fitness functions directly, they did prove helpful when choosing the final run parameters.

GMax-bio Genetic Programming Models. The protocol adopted for all GMax-bio models was to use 50% of the data as a holdout set for validation (Figure 2). This validation set was scrutinized in order to guard against overfitting. Generally, 5−10 runs, each of 500 generations, were performed using all input parameters. Parameter frequencies from these runs were then pooled and used to select a subset of the most significant parameters upon which a further set of runs were carried out (usually until no further improvement in the fitness of the model was observed for 100 generations or so). When overfitting was observed, the model was trimmed to the simplest in terms of model (tree) size, number of generations, or both to the point that preserved the best model classification errors on both the training data and the validation data.

RESULTS

Figure 3 shows typical GC×GC-TOF-MS 3D chromatograms from (a) experiment 7, one of the better runs during the first 50 experiments, and (b) from the finally optimized set of conditions. It is evident that a considerably larger number of peaks are present when the optimized conditions are used and that some of this improvement is due to improved separation in the second dimension.

Figure 4a shows the evolution of the peak number during the experimental progression. A number of features are apparent from the data. First, the improvement in the median peak number is considerable, from 469 to 3154 when the first 50 and last 50 optimization runs are compared. The final set of conditions (Table 4), carried out in sextuplicate, yield a mean of 4334 ± 155 peaks, amounting to a CV of 3.6%. This variation is mainly due to peaks with the lowest signal/noise ratio becoming differentially observable at the limit of detection margin, with improved conditions providing higher responses (or concentration in the final peak). There is also a potential contribution via inaccurate identification of peaks by library searching of mass spectra containing high proportions of noise peaks, which influence mass spectral library searches and results in duplicate peaks not being assigned as duplicate peaks. That the separation in the second dimension needs to be optimized rather than maximized is illustrated by the fact that (above a threshold) it is in fact the lower nearest-neighbor distances that are optimal (Figure 4), since when there is a restricted overall time in the second dimension, methods that keep the distances between peaks small allow more of the peaks to be observed without overlap.

The output data of the GCTofControl program give fitness values for all experiments run, and the analyst is therefore able to select appropriate tradeoffs between different fitnesses, taking into account practical experimental considerations as well as the importance of the various fitnesses for the final application. Optimum parameter values are rounded taking into consideration the spread of fitness values as well as the position of the chosen optimum. The final set of optimized conditions is shown in Table

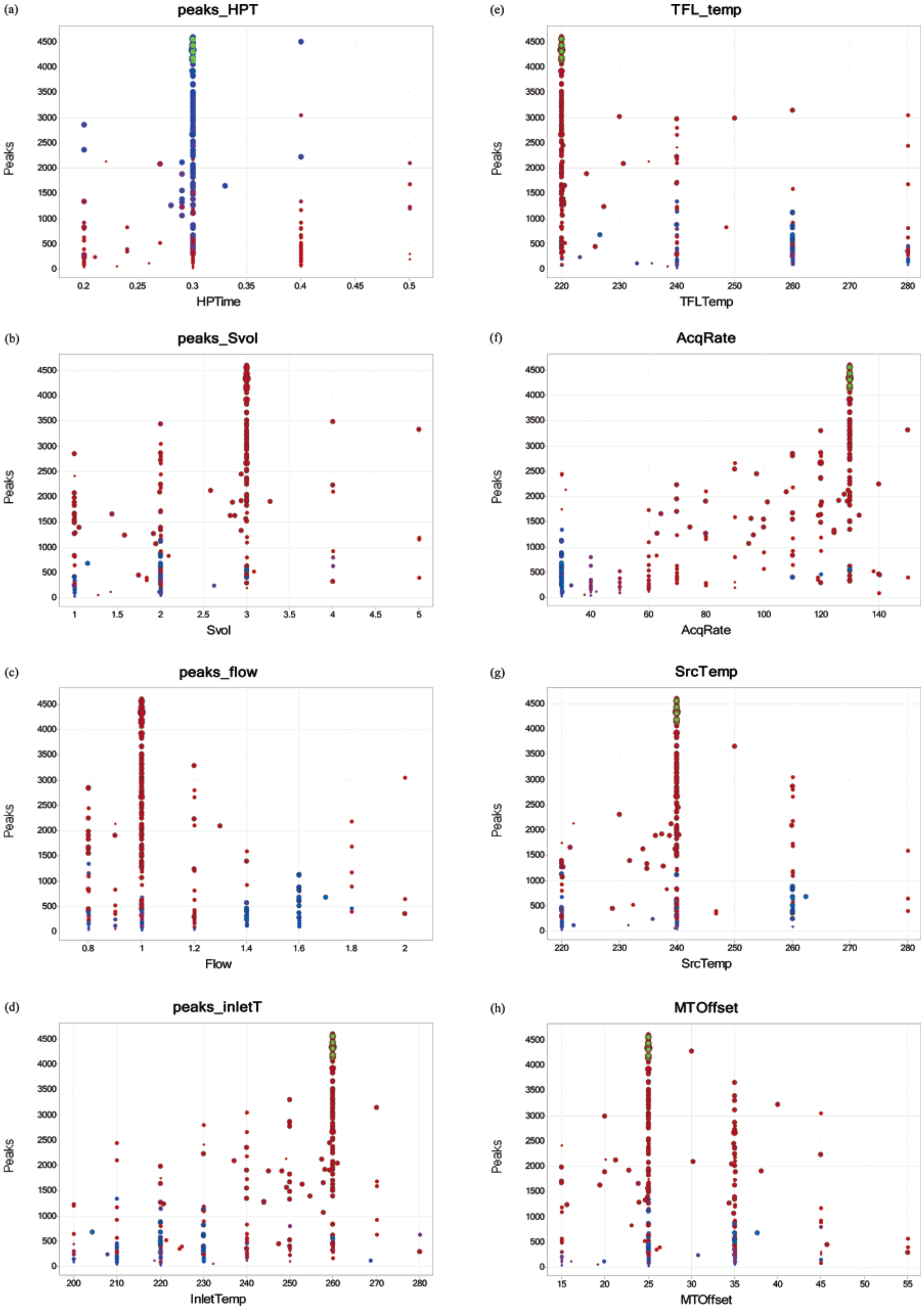


Figure 6. Effect of individual instrumental parameters on the number of peaks observed during the optimization of GC×GC-TOF MS. (a) Hot pulse time, (b) flow rate, (c) split ratio, (d) inlet temperature, (e) transfer line temperature, (f) acquisition rate, (g) source temperature, and (h) MT offset. Conditions and meaning of symbols as in Figures 4 and 5.

4. Note that, after the initial optimization of 300 experiments, a small series of experiments was performed to ensure that late-eluting peaks (on column 1 or 2) were indeed observed with the optimized conditions. In these experiments, it was discovered that cholesterol elutes late in both retention times 1 and 2 and therefore the oven 1 and 2 final temperature duration was extended from

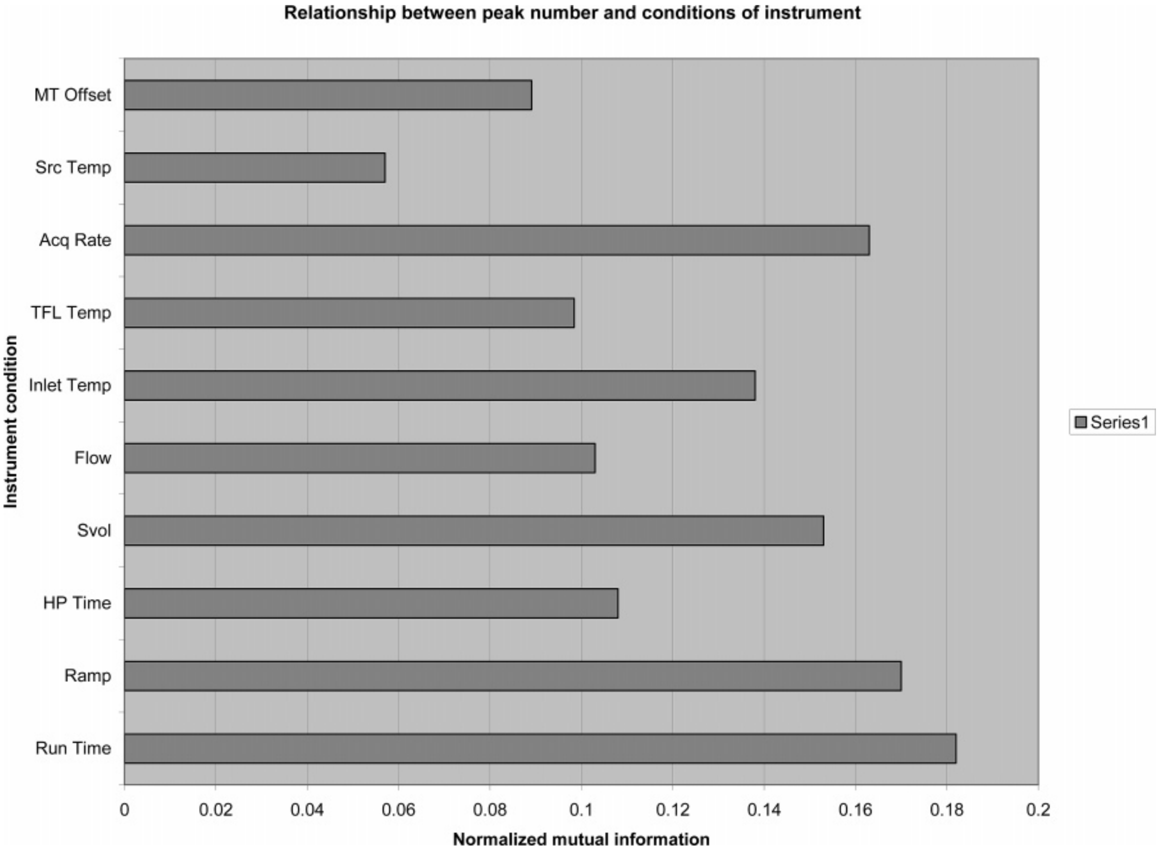


Figure 7. Normalized mutual information between the peak number and 10 conditions of the instrument's configuration.

Table 5. Comparison of Peak Numbers Observed Using Optimized Versions of 1D[35] and 2D (This Paper) GC-TOF-MS[a]

analytical technique	raw peak count from deconvolution	operator defined true peak count	metabolite peaks detected in nonpooled serum samples
GC-TOF-MS	1208	951	mean, 226
GC×GC-TOF-MS	4334	1787	mean, 694

[a] The raw peak count, for the optimized set of analytical conditions is the number of peaks reported by the Leco ChromaTof software, without further data processing. The operator-defined true metabolite peaks are those peaks which a human expert (W.B.D.) believes to be real metabolite peaks and not impurities or peaks of S/N < 5 but that are still reported. Metabolite peaks detected in nonpooled samples are the typical number of peaks detected in individual serum samples for biomarker studies.

the optimized value of 5 to 10 min and the second dimension time of 6 s was maintained.

As previously for 1D GC-TOF-MS experiments,[35] there is a significant tendency of a lengthened run time to improve the number of peaks (Figure 4a,b). The chief cause of this was the lower temperature ramp speed in the longer runs (Figure 5). This shows the greater requirement of chromatographic resolution of metabolite peaks and that one should not rely on the instrument vendor's deconvolution software package for reliable and maximal biologically relevant data to be produced.

An indication that the optimization had indeed been successful was obtained by looking at the effect of individual parameters on the number of peaks. Thus, Figure 6a shows the effect of the hot pulse time (the time that hot nitrogen jets are in operation in each modulation step) on the number of peaks, suggesting that the more successful runs had indeed found an optimum within the range studied and that the later runs were effectively optimizing or effecting a local search. Similar statements are true for the other instrumental parameters, of which another seven are illustrated in Figure 6b−h. These parameters are of interest for understanding the operation of GC×GC instrumentation for other applications. Sample volume is of interest as the thin film thickness of column 2 requires a low sample volume to ensure the column stationary phase is not overloaded and result in an increase in peak width and therefore reduces chromatographic resolution. In this optimization, a maximum on-column volume of 0.5 μL was employed to ensure overloading of column 2 was not observed. As for the GC-TOF-MS optimization previously reported,[35] higher sample volumes ensure that lower concentration metabolites are still detected. Of interest is that low sample volume/low split ratios are preferentially chosen to high sample volume/high split ratios, even though similar on-column sample volumes are observed for both.

The inlet and source temperatures are important to ensure that metabolites, more specifically their oximated and trimethylsilyl derivatives, are sufficiently volatile to traverse heated regions within the analytical system while also ensuring that temperatures are not too high as to degrade derivatized molecules thermally.

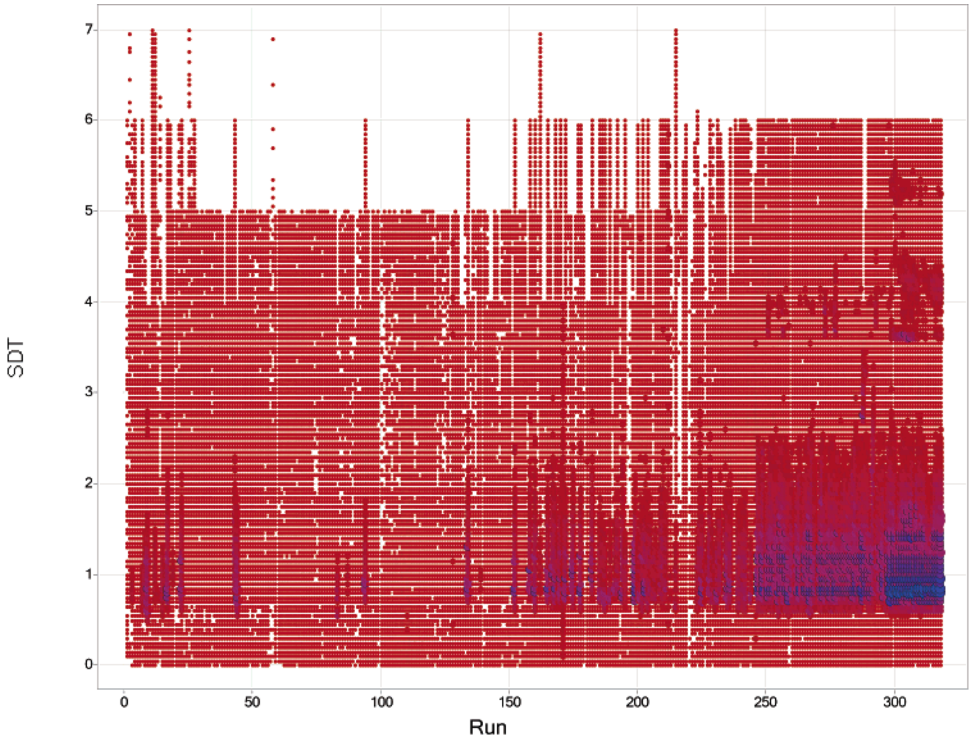


Figure 8. Illustration of the increase (and distribution) of the second dimensional separation as a function of run time. The symbols encode the local density of peaks, while the color and symbol size encode the total peak number.

The inlet and source temperatures of 260 and 240 °C, respectively, provide the transfer of the greatest number of metabolite peaks will minimizing metabolite peak thermal degradation.

Of specific interest for GC×GC applications is the transfer line temperature, especially as short column 2 lengths are used. The percentage of column 2 that lies outside the column 2 oven is much greater for GC×GC than for GC applications; in this example, one-third of column 2 (50 cm) is located in the transfer line and not in the oven, with both operating at different temperatures (especially column 2 whose temperature is ramped during an analytical run). It was found that a transfer line temperature of 220 °C was optimal when compared to higher temperatures, presumably because lower temperatures provide better chromatographic resolution for column 2.

To give a further indication of the relationship between the peak number and the settings of the instrument, the normalized mutual information between peak number and each of the 10 conditions considered in Figures 4−6 has been calculated and is displayed in Figure 7. The mutual information between two variables X and Y is an information-theoretic measure of their mutual dependence;[57−59] it is a symmetric measure, which quantifies the difference between the joint distribution of X and Y and their marginal distributions. Here, we compute the normalized mutual information from the data collected during the optimization, using NMI = $[2\mathrm{MI}(X,Y)/(H(X) + H(Y))]^{1/2}$, where MI($X$,$Y$)-is the mutual information of X and Y, given by MI(X,Y) = $\sum_n\sum_n p(x \wedge y)\log_2 p(x \wedge y)/p(x)p(y)$ and $H(X)$ and $H(Y)$ are the entropy of X and Y, respectively. To compute the mutual information and entropies, the data were binned into $n = 10$ bins and the sample probabilities of the X and Y variables falling into each bin are calculated.

Figure 7 shows that all 10 of the conditions considered have an influence on the number of peaks and that run time, ramp, and acquisition rate have a particularly large effect. This is of particular importance, showing that all conditions do indeed have a significant influence on the analytical result and therefore need to be part of the heuristic search.

Overall, these peak numbers compare very favorably with those seen in the 1D optimized GC-TOF (Table 5).

It is reasonable next to enquire as to where these extra peaks come from. One reason for the greater number of peaks is the improved S/N achievable with GC×GC, when compared to GC strategies. Modulation provides narrow peak widths in the second dimension (typically 0.2 s compared to 3 s for GC), which improves S/N. Also metabolite peaks are chromatographically separated from the chemical background (solvent and derivatization peaks observed at low second dimension retention times), which also improves S/N compared to GC applications. The second possibility is that these "extra" peaks are essentially peaks of low S/N that were incompletely deconvolved from (or completely hidden by) larger ones when the 1D separation was inadequate to discriminate them. Given that the deconvolution algorithm in the Leco software (and most other deconvolution software packages) relies on features in the mass spectra, it cannot, for completely overlapping peaks, discriminate small peaks in terms of whether they come as less common fragments of a more abundant peak rather than major fragments of a less abundant peak. Complementary to Figure 3, Figure 8 gives an indication of how more peaks do indeed appear in the second

[57] Shannon, C. E.; Weaver, W. *The mathematical theory of communication*; University of Illinois Press: Urbana, IL, 1949.

[58] Battiti, R. *IEEE Trans. Neural Networks* **1994**, *5*, 537−550.

[59] Broomhead, D. S.; Sidorov, N. *Nonlinearity* **2004**, *17*, 2203−2223.

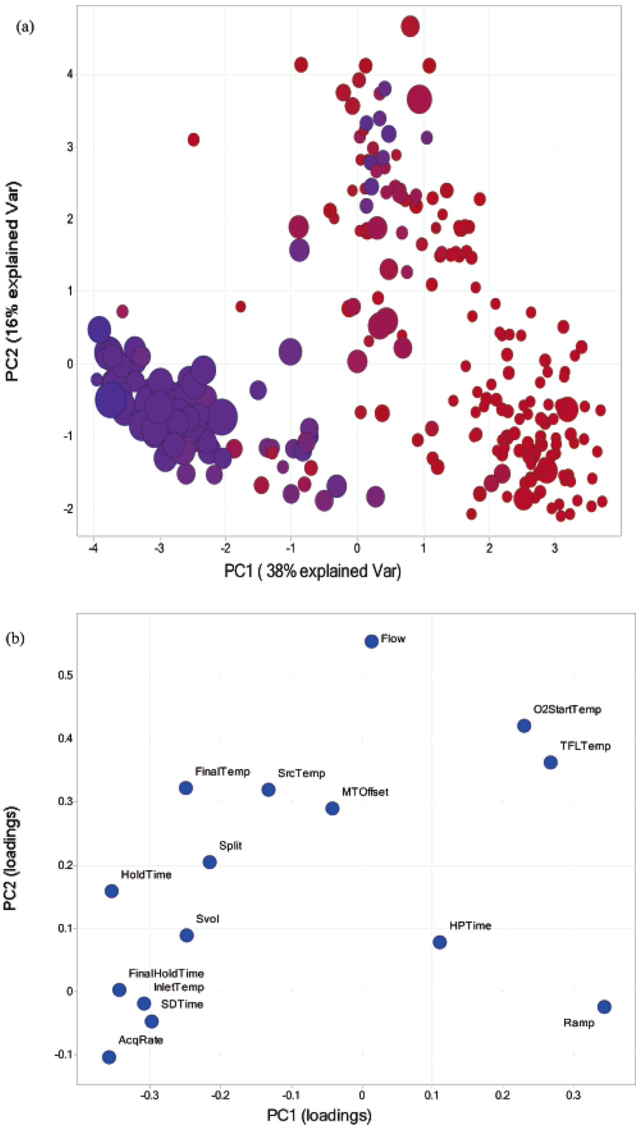


Figure 9. "Landscape" of the GC×GC optimization, with the landscape encoded as principal components of the variance of the parameters for purposes of visualization. (a) Scores, showing the peak numbers encoded by symbol size and the run time by color (low red, high blue). (b) Loadings. The contribution of the each of the parameters studied to the variance in the model.

dimension as the run number increases and thus must have been "hidden" during the earlier runs.

It is also of interest to seek to understand the overall optimization landscape, which in the case of the electrospray mass spectrometry optimization[26] appeared to be very epistatic indeed (in the sense that the optimal value of one variable depended on the values of other variables, although that for 1D GC-TOF was much less so).[35] It might be the case, for instance, that the extra degrees of freedom offered by the second dimension could change the landscape over that in 1D GC significantly, making it even less epistatic. While an inspection of the data of Figure 6 could be consistent with that view, we recognize that we used a different algorithm here, and a straight comparison is inappropriate. Nevertheless, we illustrate the landscape in Figure 9, using as before the first two principal components of the variance of the parameters to reduce their dimensionality for purposes of visualization.

The plot of PC1 versus PC2 in Figure 9a clearly shows three separate, almost orthogonal, clusters. The bottom left cluster reflects the experimental conditions producing both the maximum number of peaks and also the maximum run times. The top cluster shows a good spread of peak detection, with a couple of reasonably high values of peak number, while keeping the run time relatively low. The third cluster (bottom right) contains mainly low peak numbers combined with short run times. The cluster distribution is further explained by the loadings plot (Figure 9b). Cluster 1 is significantly influenced by time-based parameters (AcqRate, SD-time, FinalHoldTime, HoldTime), while cluster 2 is significantly influenced by temperature components and flow, and cluster 3 is significantly influenced by ramp and HPtime. Together with the earlier data, especially those in Figure 7, these plots illustrate not only how the system improved but why.

DISCUSSION

The history of biochemistry is replete with important advances that have been occasioned by the discovery of novel metabolites that, as well as their intrinsic scientific interest, might also have significance in applied work and in medicine. As with proteomics,[60] the large dynamic range of the human serum metabolome[61] means that inadequate separations will cause substances present at low concentrations to comigrate with components present in much larger concentrations, thereby obscuring both their detection and their identification. Even with mass spectral information this can make their deconvolution extremely challenging. Consequently, both prefractionation and improved separations are among the better strategies for increasing the number of metabolites that may be detected in metabolomics experiments. Our preference where possible is for the latter, which is a more appropriate strategy for the high-throughput approaches that are required (for statistical reasons if no other[62−65]) in metabolomics.

Exhaustive search of the possible combinations of chromatographic conditions that might be used is out of the question, and so heuristic methods are appropriate. Even then, potentially hundreds of experiments must still be performed, and automation then becomes a very desirable approach. In the present work, we extended our earlier closed-loop strategy[35] to 2D GC, increasing the number of "raw" peaks to over 4000 and the number of discernible metabolites to ~1800. In addition, we used a different multiobjective optimization algorithm,[46] which is considered highly efficient for continuous functions.

The serum employed was a commercially available pooled sample from 17 individuals. The variability of the human metabolome has been recognized for a long time[66] and previously detailed, for instance, in terms of factors such as age,[67] gender,[68,69] diet

[60] Anderson, N. L.; Anderson, N. G. *Mol. Cell. Proteomics* **2002**, *1*, 845−867.
[61] Kell, D. B. *Curr. Opin. Microbiol.* **2004**, *7*, 296−307.
[62] Ioannidis, J. P.; Trikalinos, T. A.; Ntzani, E. E.; Contopoulos-Ioannidis, D. G. *Lancet* **2003**, *361*, 567−571.
[63] Wacholder, S.; Chanock, S.; Garcia-Closas, M.; El, Ghormli, L.; Rothman, N. *J. Natl. Cancer Inst.* **2004**, *96*, 434−442.
[64] Ioannidis, J. P. *PLoS Med.* **2005**, *2*, e124.
[65] Ein-Dor, L.; Zuk, O.; Domany, E. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 5923−5928.
[66] Williams, R. J. *Biochemical Individuality*; John Wiley: New York, 1956.
[67] Williams, R. E.; Lenz, E. M.; Rantalainen, M.; Willson, I. D. *Mol. Biosyst.* **2006**, *2*, 193−202.
[68] Stanley, E. G.; Bailey, N. J. C.; Bollard, M. E.; Haselden, J. N.; Waterfield, C. J.; Holmes, E.; Nicholson, J. K. *Anal. Biochem.* **2005**, *343*, 195−202.

and culture,[70] and health/disease status of the subject[71,72] being likely to influence the composition of a particular metabolome. This has also been shown in studies within the authors' laboratory, where typically 500−800 metabolite peaks are detected in serum obtained from specific individuals, compared to the 1800 detected in pooled serum as employed in this optimization study. It can be expected that as greater numbers of samples are studied more novel metabolite peaks will be detected. The primary objective of the HUSERMET project (http://www.husermet.org/) is to assess the variability of the human metabolome, both in composition and in concentration, by analysis of serum obtained from more than 5000 individuals. Currently, over 2600 metabolites have been estimated by genome-scale reconstruction models (Palsson, personal communication, cf. ref 73), while the study of other biofluids such as urine will extend the library of metabolites further, including those derived from gut microflora.[34]

Of the nearly 1800 metabolite peaks detected, we have currently identified only 188 metabolites by mass spectral library searches using the mass spectral libraries defined above (similarity >750), showing the major need to identify these metabolite peaks by running authentic standards and thereby including more known (or estimated) metabolites in mass spectral libraries. It should be noted that many metabolites produce multiple derivatization products, and therefore, in this study more than 350 peaks have been assigned an identification, and it is this that is equivalent to 188 metabolites. Current work is ongoing to identify metabolites definitively in this way by the compilation of a mass spectral/retention index library for the optimized set of conditions reported, including those for metabolites not currently available in academic or commercially available mass spectral libraries.

Based on the heuristic analysis of the separations landscape in this system, it does not appear likely that major improvements in the number of peaks will now come from improving the chromatographic separations per se. However, we recognize that much greater resolution in the mass spectral dimension is possible, including the use of exact mass techniques,[74] and this is evidently an important strategy both for further improving the number of metabolites that may be detected and for assisting their identification. Another approach that we are pursuing is to seek to identify in the 2D GC×GC-TOF-MS data all the metabolites that we would expect from the known biochemistry to be present in human serum. If the deconvolution problem is reduced to deconvolving known coeluting substances from each other, this will permit the application of different and more powerful chemometric techniques (cf. refs 53 and 75−81).

In conclusion, by using an advanced, closed-loop optimization method, we have demonstrated a substantial improvement in the number of human serum metabolites that may be detected reliably using GC×GC-TOF-MS. Understanding their nature and distribution between individuals under different conditions is now possible, where the existence of a suitable data model and database[55] will make this task much easier.

ACKNOWLEDGMENT

We thank the BBSRC and MRC, partly under the terms of the LINK scheme in Applied Genomics (the HUSERMET project), for financial support, as well as the EPSRC, RSC, and BHF. We also thank AstraZeneca and GlaxoSmithKline for support of the HUSERMET project. We thank Nick Bukowski (Anatune), Roy Goodacre (Manchester), John Haselden (GSK), Andy Nicholls (GSK), and Ian Wilson (AZ) for useful discussions. S.O'H., W.B.D., and J.D.K. contributed equally to this work.

Received for review August 3, 2006. Accepted October 25, 2006.

AC061443+

(69) Kochhar, S.; Jacobs, D. M.; Ramadan, Z.; Berruex, F.; Fuerhoz, A.; Fay, L. B. *Anal. Biochem.* **2006**, *352*, 274−281.

(70) Lenz, E. M.; Bright, J.; Wilson, I. D.; Hughes, A.; Morrisson, J.; Lindberg, H.; Lockton, A. *J. Pharm. Biomed. Anal.* **2004**, *36*, 841−849.

(71) Kenny, L. C.; Dunn, W. B.; Ellis, D. I.; Myers, J.; Baker, P. N.; The GOPEC Consortium; Kell, D. B. *Metabolomics* **2005**, *1*, 227−234; online DOI: 210.1007/s11306−11005.-10003−11301.

(72) Underwood, B. R.; Broadhurst, D.; Dunn, W. B.; Ellis, D. I.; Michell, A. W.; Vacher, C.; Mosedale, D. B.; Kell, D. B.; Barker, R.; Grainger, D. J.; Rubinsztein, D. C. *Brain* **2006**, *129*, 877−886.

(73) Romero, P.; Wagg, J.; Green, M. L.; Kaiser, D.; Krummenacker, M.; Karp, P. D. *Genome Biol.* **2005**, *6*, R2.

(74) Williams, R.; Lenz, E. M.; Wilson, A. J.; Granger, J.; Wilson, I. D.; Major, H.; Stumpf, C.; Plumb, R. *Mol. Biosyst.* **2006**, *2*, 174−183.

(75) Brereton, R. G.; Dunkerley, S. *Analyst* **1999**, *124*, 705−711.

(76) Demir, C.; Hindmarch, P.; Brereton, R. G. *Analyst* **2000**, *125*, 287−292.

(77) Woodward, A. M.; Rowland, J. J.; Kell, D. B. *Analyst* **2004**, *129*, 542−552.

(78) Sinha, A. E.; Hope, J. L.; Prazen, B. J.; Fraga, C. G.; Nilsson, E. J.; Synovec, R. E. *J. Chromatogr., A* **2004**, *1056*, 145−154.

(79) Smilde, A.; Bro, R.; Geladi, P. *Multi-way analysis: applications in the chemical sciences*; Wiley: New York, 2004.

(80) A, J.; Trygg, J.; Gullberg, J.; Johansson, A. I.; Jonsson, P.; Antti, H.; Marklund, S. L.; Moritz, T. *Anal. Chem.* **2005**, *77*, 8086−8094.

(81) Katajamaa, M.; Miettinen, J.; Orešič, M. *Bioinformatics* **2006**, *22*, 634−636.