Explanatory Optimization of Protein Mass Spectrometry via Genetic Search

Seetharaman Vaidyanathan,[†,‡] David I. Broadhurst,[‡] Douglas B. Kell,*[,†,‡] and Royston Goodacre[†,‡]

Department of Chemistry, UMIST, P.O. Box 88, Sackville Street, Manchester M60 1QD, U.K., and Institute of Biological Sciences, University of Wales, Aberystwyth, Ceredigion SY23 3DD, U.K.

Optimizing experimental conditions for the effective analysis of intact proteins by mass spectrometry is challenging, as many analytical factors influence the spectral quality, often in very different ways for different proteins and especially with complex protein mixtures. We show that genetic search methods are highly effective in this kind of optimization and that it was possible in 6 generations with a total of $<$500 experiments out of some 10^{14} to find good combinations of experimental variables (electrospray ionization mass spectral settings) that would not have been detected by optimizing each variable alone (i.e., the search space is epistatic). Moreover, by inspecting the evolution of the variables to be optimized using genetic programming, we discovered an important relationship between two of the mass spectrometer settings that accounts for much of this success. Specifically, the conditions that were evolved included very low values of skimmer 1 voltage (the sample cone) and a skimmer 2 voltage (extraction cone) above a threshold that would nevertheless minimize the potential difference between the sample and extraction skimmers. The discovery of this relationship demonstrates the hypothesis-generating ability of genetic search in optimization processes where the size of the search space means that little or no a priori knowledge of the optimal conditions is available.

There is much current interest in the exploitation of soft ionization mass spectrometries in proteomic studies.[1,2] In particular, mass spectral analysis of"intact" proteins in complex mixtures using electrospray ionization mass spectrometry (ESI-MS) is gaining momentum,[3−6] thanks to improvements in mass spectral resolution and sensitivity[7] and the relevance of "top-down" strategies for mass spectral characterization of proteomes.[8,9] High-throughput proteomics increasingly demands the capability to identify and characterize as many proteins as possible from complex mixtures with minimal recourse to cleanup or separation stages prior to MS. Thus, any strategy that maximizes the coverage of proteins with minimal operational stages within an analysis is highly desirable. In this context, the direct analysis of intact proteins from mixtures by mass spectrometry is attractive.

It is known that factors such as pH,[10] ionic strength,[11] the solvent used to dissolve the protein,[12,13] and instrumental settings[14−16] influence the effective ionization and detection of proteins, even when individual proteins are analyzed in isolation. Successful strategies for the analysis of intact proteins in complex mixtures therefore impose more strenuous requirements on optimizing the experimental conditions for their analysis. In particular, mass spectrometer conditions that are effective in ionizing a particular peptide or protein may be very poor for other proteins. However, this does not mean that more "universal" and effective mass spectrometric conditions might not be found. The problem is that the number of possible conditions we might try is absolutely enormous, since for *n* mass spectrometer parameters that may be varied, each of which may take *m* values, the number of combinations (known as the "search space") is m^n. Such problems are called combinatorial optimization problems.[17] These are typically NP-complete problems (problems not currently solvable in a deterministic polynomial time),[18] which scale very poorly (exponentially) with *n*, such that trying every combination ("exhaustive search"), even for modest values of both *m* and *n*, is

* Corresponding author. E-mail: dbk@umist.ac.uk. Tel: +44 (0)161 200 4492. Fax: +44 (0)161 200 4556.

[†] UMIST.

[‡] University of Wales.

(1) Mann, M.; Hendrickson, R. C.; Pandey, A. *Annu. Rev. Biochem.* **2001**, *70*, 437−473.
(2) Aebersold, R.; Mann, M. *Nature* **2003**, *422*, 198−207.
(3) Li, W. Q.; Hendrickson, C. L.; Emmett, M. R.; Marshall, A. G. *Anal. Chem.* **1999**, *71*, 4397−4402.
(4) Rostom, A. A.; Fucini, P.; Benjamin, D. R.; Juenemann, R.; Nierhaus, K. H.; Hartl, F. U.; Dobson, C. M.; Robinson, C. V. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5185−5190.
(5) Lee, S. W.; Berger, S. J.; Martinovic, S.; Pasa-Tolic, L.; Anderson, G. A.; Shen, Y. F.; Zhao, R.; Smith, R. D. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 5942−5947.
(6) Hayter, J. R.; Robertson, D. H. L.; Gaskell, S. J.; Beynon, R. J. *Mol. Cell. Proteomics* **2003**, *2*, 85−95.
(7) Smith, R. D. *Trends Biotechnol.* **2002**, *20*, S3−S7.
(8) McLafferty, F. W. *Int. J. Mass Spectrom.* **2001**, *212*, 81−87.
(9) Reid, G. E.; McLuckey, S. A. *J. Mass Spectrom.* **2002**, *37*, 663−675.
(10) Mirza, U. A.; Chait, B. T. *Anal. Chem.* **1994**, *66*, 2898−2904.
(11) Wang, G. D.; Cole, R. B. *Anal. Chem.* **1994**, *66*, 3702−3708.
(12) Iavarone, A. T.; Jurchen, J. C.; Williams, E. R. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 976−985.
(13) Loo, J. A. *Int. J. Mass Spectrom.* **2000**, *200*, 175−186.
(14) Ashton, D. S.; Beddel, C. R.; Cooper, D. J.; Green, B. N.; Oliver, R. W. A. *Org. Mass Spectrom.* **1993**, *28*, 721−728.
(15) Hunt, S. M.; Sheil, M. M.; Belov, M.; Derrick, P. J. *Anal. Chem.* **1998**, *70*, 1812−1822.
(16) Oberacher, H.; Walcher, W.; Huber, C. G. *J. Mass Spectrom.* **2003**, *38*, 108−116.
(17) Ausiello, G.; Crescenzi, P.; Gambosi, G.; Kann, V.; Marchetti-Spaccamela, A.; Protasi, M. *Complexity and Approximation. Combinatorial optimization problems and their approximability properties*; Springer-Verlag: Berlin, 1999.
(18) Garey, M. R.; Johnson, D. S. *Computers and Intractability: A Guide to the Theory of NP Completeness*; W. H. Freeman and Co.: San Francisco, CA., 1979.

10.1021/ac034669a CCC: $25.00 © 2003 American Chemical Society
Published on Web 10/25/2003

simply impossible. For instance, with 10 variables, each of which can take just 5 values, the search space (5^{10}) is ~9.8 million possible experiments. Although knowledge of ion physics can help in minimizing possible combinations of parameters that would be relevant, the resultant optimization would result in biased observations on the variable search space, and there is always the possibility of losing valuable information. Therefore, a systematic approach to the optimization exercise is called for.

In consequence, so-called heuristic methods,[19,20] which seek solutions that approach an optimum but cannot be guaranteed to find it, are used to navigate these huge search spaces to find regions that are optimal for some desired property. Among heuristic methods, evolutionary computing techniques are popular.[19−22] Evolutionary or genetic search methods encode candidate solutions to a problem in the form of a population of "individuals" or "chromosomes", whose "fitness" may be evaluated against some desired property. According to a generalized algorithm based in part on the fitness of the individuals, some are selected to produce "offspring" via mutation (one parent only) or recombination (using two or more parents), whose fitness may be further evaluated, and the cycle is continued until a desired stopping criterion is met. Such techniques have enjoyed many successes in the effective exploration of combinatorial search spaces.[19−22] In addition, it has been argued that the purely data-driven, evolutionary exploration of large and complex search spaces can itself generate new scientific and technical knowledge,[22−26] and it was of interest to explore this view in the present problem domain.

We note that combinatorial optimization methods such as those based on evolutionary algorithms (EAs) are *especially* well suited to problems in which the variables depend on each other (otherwise one could merely use linear programming methods or classical experimental design). We here show that genetic search methods are highly effective in improving the quality of the mass spectra in complex mixtures of proteins and thereby discover an important relationship between two of the mass spectrometer variables which accounts for much of this success.

MATERIALS AND METHODS

Chemicals. Acetonitrile (HPLC grade) and water (HPLC grade) were obtained from Fisher Scientific (Loughborough, U.K.). Formic acid (FA) and five proteins, viz. insulin (bovine pancreas), ubiquitin (bovine erythrocytes), cytochrome *c* (equine heart), lysozyme (hen egg white), and myoglobin (equine skeletal muscle) were purchased from Sigma (Dorset, U.K.). Stock solutions (30 μM) of the individual proteins were prepared in 0.1% FA. An equimolar mixture of the five proteins, diluted 1:1 with acetonitrile (final concentration of each protein, 1 μM), was used for the analysis.

Table 1. Variables and Their Range

code	variables	range
V1	sample flow rate (μL/min)	20−500
V2	desolvation gas flow rate (L/h)	150−500
V3	nebulizer gas flow rate (L/h)	10−20
V4	source tempearture (°C)	40−150
V5	desolvation temperature (°C)	100−400
V6	capillary voltage (V)	1500−3500
V7	skimmer 1 (sample cone) voltage (V)	10−150
V8	skimmer 2 (extraction cone) voltage (V)	0−10
V9	transport hexapole voltage (V)	0−20
V10	differential pumping aperture voltage (V)	0−20
V11	acceleration lens voltage (V)	0−200
V12	focus voltage (V)	0−200
V13	prefilter voltage (V)	5−15
V14	MCP detector voltage (V)	2300−2700

Mass Spectrometry. ESI-MS was performed in the flow injection mode[27−29] using a Q-TOF 1.5 mass spectrometer (equipped with a Z-spray), supplied by Micromass Ltd. (Manchester, U.K.). The TOF analyzer in the mass spectrometer is arranged in an orthogonal configuration. Spectra were acquired in the positive (ES+) ion mode, between m/z 300 and 2000, with relevant instrumental parameters set in the ranges given in Table 1. Spectra were acquired every 3 s, and acquisitions over the duration of the injected volume (ranging from 2 to 10 min) were combined to give the mass spectrum. Myoglobin (M_r = 16 950 Da) was used to tune the instrument. The protein mixture was introduced into the mass spectrometer using the autosampler of a Waters 2790 liquid chromatography separation unit (without the column). A mobile liquid phase of 50% aqueous acetonitrile containing 10 mM FA was injected into the mass spectrometer at a flow rate ranging from 20 to 500 μL min^{-1}, and an aliquot (30 μL) of the sample was loaded from a 300-μL 96-well microtiter plate, maintained at 8 °C, directly into the flow stream and on into the ionization source of the mass spectrometer. To ensure that any substantive instrumental drift could be taken into account, the experiments were carried out over a 6-month period, although the acquisition of data for each generation was typically carried out over 2 days. The parameter settings were set by the user using the manufacturer's software.

Data Processing. The spectral data were normalized to total ion counts and exported from MassLynx (Micromass) to Matlab (Maths Works), at 0.1 amu resolution. The normalized spectra were then analyzed using a routine written in Matlab to match peak positions with respect to those ideally expected for a mixture of the five proteins and give out various parameters that relate to the matched peaks with respect to the number of peaks, the percentage of expected peaks present for each protein, the signal-

[19] Corne, D., Dorigo, M., Glover, F., Eds. *New ideas in optimization*; McGraw-Hill: London, 1999.

[20] Michalewicz, Z.; Fogel, D. B. *How to solve it: modern heuristics*; Springer-Verlag: Heidelberg, 2000.

[21] Bäck, T., Fogel, D. B., Michalewicz, Z., Eds. *Handbook of evolutionary computation.*; IOPPublishing/Oxford University Press: Oxford, 1997.

[22] Goldberg, D. E. *The design of innovation: lessons from and for competent genetic algorithms*; Kluwer: Boston, 2002.

[23] Gillies, D. *Artificial intelligence and scientific method*; Oxford University Press: Oxford, 1996.

[24] Koza, J. R.; Bennett, F. H.; Keane, M. A.; Andre, D. *Genetic Programming III: Darwinian Invention and Problem Solving*; Morgan Kaufmann: San Francisco, 1999.

[25] Kauffman, S.; Lobo, J.; Macready, W. G. *J. Econ. Behav. Organization* **2000**, *43*, 141−166.

[26] Koza, J. R.; Keane, M. A.; Streeter, M. J.; Mydlowec, W.; Yu, J.; Lanza, G. *Genetic programming: routine human-competitive machine intelligence*; Kluwer: New York, 2003.

[27] Vaidyanathan, S.; Kell, D. B.; Goodacre, R. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 118−128.

[28] Allen, J. K.; Davey, H. M.; Broadhurst, D.; Heald, J. K.; Rowland, J. J.; Oliver, S. G.; Kell, D. B. *Nat. Biotechnol.* **2003**, *21*, 692−696.

[29] Goodacre, R.; Vaidyanathan, S.; Bianchi, G.; Kell, D. B. *Analyst* **2002**, *127*, 1457−1462.

to-noise ratio, and the relative intensity contributions for the individual proteins. These parameters were then used to calculate a fitness function for evaluating the evolutionary algorithm (genetic search).

We set up the first generation of 180 experiments by taking random values for each of the parameters, within the specified ranges (Table 1). The position (m/z) of the charge state peaks for each protein under the ionization conditions. The protein charge-state peaks that were observed for each protein, when analyzed in isolation and under a standard set of conditions, were taken to form an "ideal" spectrum, with respect to the peak positions (m/z values). An "ideal" spectrum within m/z 600−2000 thus comprised peaks for charge states 3+ to 9+ for insulin, 5+ to 14+ for ubiquitin, 7+ to 20+ for cytochrome *c*, 8+ to 14+ for lysozyme, and 9+ to 27+ for myoglobin. Three separate properties of the protein mass spectra were considered to contribute to the fitness: (a) one that maximizes the coverage of the individual protein (charge-state) peaks in the mixture (relative to the "ideal" spectrum) − peak positions (Pk_{idx}), (b) one that maximizes evenness of signal contribution from the five proteins, or the contribution of peak intensities of the different charge states of each protein to the total spectral intensity − relative intensity contribution (I_{idx}), and (c) one that maximizes signal (charge-state peaks)-to-noise (nonprotein/protein fragment peaks) (s/n). Although we could have treated this as a multiobjective problem,[30−32] we combined these to make a combined fitness function, as follows:

$$Pk_{\mathrm{idx}} = (N_1/X_1)(N_2/X_2)(N_3/X_3)(N_4/X_4)(N_5/X_5)$$

$$I_{\mathrm{idx}} = (I_1/I)(I_2/I)(I_3/I)(I_4/I)(I_5/I)$$

$$s/n = \left(\frac{\sum I_y}{I - \sum I_y}\right)$$

$$\mathrm{fitness} = (Pk_{\mathrm{idx}})(I_{\mathrm{idx}})(s/n)$$

where N_y is the number of peaks that are observed for protein Y, X_y is that ideally expected (those in the "ideal" spectrum), I_y is the combined intensity of the observed charge states for protein Y, and I is the total spectral intensity. The product of the individual contributions was taken so as to penalize the fitness if any individual component was absent (i.e., takes a value of zero). Maximizing the fitness function would thus maximize both the evenness of contribution of each protein to the total intensity and the s/n. After each generation, we evaluated the fitness of each individual encoding the mass spectral parameters and produced a new generation, continuing for a total of six generations.

Evolutionary Algorithm. A set of experiments in which the levels of the variables were set in a random order was chosen as the first generation. The settings for the subsequent generation were defined by feeding this information to the EA, which in turn generated a set of candidates to be examined in the second generation, the results of which were then fed back to the EA for generating the next generation. This procedure was carried out in an iterative manner until the sixth generation was reached. The EA was written in-house, running under Microsoft Windows NT on an IBM-compatible PC. A variation on the simple genetic algorithm (GA) developed by Holland[33] was used. In our study, the GA used proportional selection, and two-point crossover with mutation, operating on a population of binary-encoded chromosomes, each chromosome representing *n* parameter values. The length of the chromosome is dependent on the resolution and range of each of the *n* parameters. For example, a parameter with the range 0−100 units and resolution of 5 units will be encoded as a 5-bit string, and a parameter with the range 10−15 units and a resolution of 1 unit will be encoded as a 3-bit string. Each chromosome in the GA therefore consists of a single binary string containing all the encoded parameter values (i.e., instrument settings) in a set sequence.

After initialization and the first set of fitness evaluations have been made, parents are selected to create the next generation using "fitness-proportional selection",[33] where the probability of selection is directly proportional to fitness. Child chromosomes are created by recombining two parent chromosomes using the two-point crossover strategy.[34] The crossover points are chosen randomly from anywhere along the whole chromosome length; i.e., they are not restricted to parameter boundaries. The probability of mutating a given chromosome after recombination was set to 0.2, and the probability of changing a bit from a 0 to 1 (or vice versa) once a chromosome is selected for mutation was set to 0.01. The process of selection, recombination, and mutation is repeated until a new population is produced. No two identical candidates are allowed in a given generation, and the top 10% of each generation are automatically transferred unchanged to the next generation. This is known as an elitist strategy and is guaranteed to converge to an optimal solution.[35] If a chromosome is created that is identical to one "evaluated" previously in the GA run and therefore already has a known fitness value (this includes the top 10% from the previous generation), this chromosome is ignored and the selection is repeated.

In addition, to assess the contribution of different experimental variables to the fitness, we performed genetic programming analysis (see refs 36−38) of the data set using the program Gmax-bio (Aber Genomic Computing, Aberystwyth, Wales), with default settings proposed by the manufacturer.

RESULTS

Electrospray ionization of a protein produces a "comb"-like spectrum of peaks with different mass/charge (m/z) ratios,[39] resulting from a distribution of charge states. Figure 1A−E shows

[30] Zitzler, E. *Evolutionary algorithms for multiobjective optimization: methods and applications*; Shaker Verlag: Aachen, 1999.

[31] Knowles, J. D.; Corne, D. W. *Evol. Comput.* **2000**, *8*, 149−172.

[32] Deb, K. *Multi-objective optimization using evolutionary algorithms*; Wiley: New York, 2001.

[33] Holland, J. H. *Adaptation in Natural and Artificial Systems*; The University of Michigan Press: Ann Arbor, MI, 1975.

[34] Goldberg, D. E. *Genetic Algorithms in search, optimization and machine learning*; Addison-Wesley: Reading, MA, 1989.

[35] Rudolph, G. *Convergence properties of evolutionary algorithms*; Verlag Dr Kovac: Hamburg, 1997.

[36] Koza, J. R. *Genetic programming: on the programming of computers by means of natural selection*; MIT Press: Cambridge, MA, 1992.

[37] Langdon, W. B. *Genetic programming and data structures: genetic programming + data structures = automatic programming!*; Kluwer: Boston, 1998.

[38] Kell, D. B. *Trends Genet.* **2002**, *18*, 555−559.

[39] Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. *Science* **1989**, *246*, 64−71.

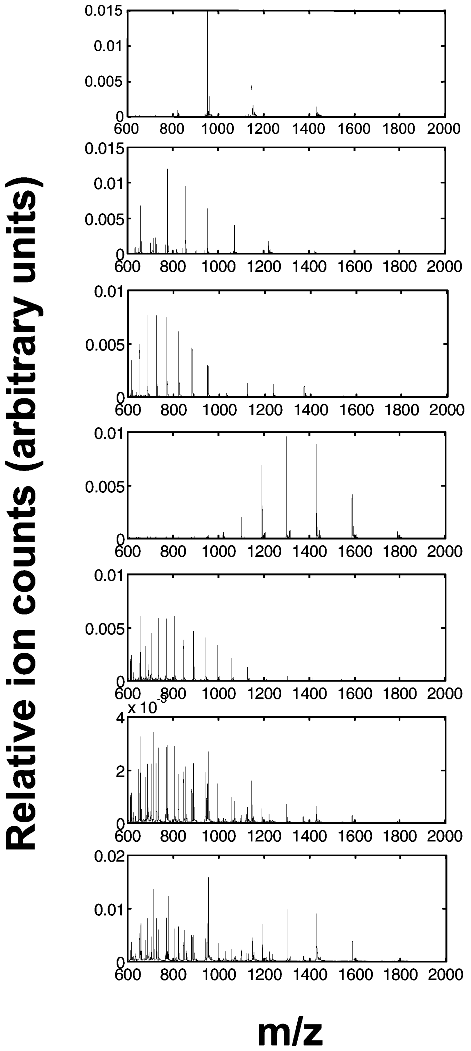

Figure 1. Positive-ion ESI-TOF mass spectra of five proteins: (A) insulin (5.7 kDa), (B) ubiquitin (8.6 kDa), (C) cytochrome *c* (12.3 kDa), (D) lysozyme (14.3 kDa), and (E) myoglobin (16.9 kDa), (F) a spectrum of an equimolar mixture of the five proteins, and (G) their combined theoretical mixture spectrum, all obtained under one set of instrumental conditions. The spectra were acquired with a flow rate of 120 μL/min. The MS was operated at a capillary voltage of 3200 V, the extraction cone voltage was 5 V, and the sample cone voltage was 35 V. The source and desolvation temperatures were 80 and 250 °C, respectively, while the desolvation and nebulizer gas flow rates were 350 and 20 L/h, respectively.

the mass spectra of the five proteins, electrosprayed individually, using one set of standard conditions. The mass spectrum of an equimolar mixture of the proteins obtained under the same conditions is shown in Figure 1F, while Figure 1G is an “ideal” mixture spectrum that is constructed by summing the spectra of the individual proteins. As expected, both the nature (shape of the comb) and the effectiveness of ionization (relative ion counts) differ for the individual proteins. It can also be seen that the mixture spectrum is quite different from the combined spectrum of the mixture components. The most obvious difference is seen with the lysozyme peaks. The relative signal from this protein seems to be influenced by the presence of the other proteins in the sample matrix. Such observations have been noted by other investigators and have been argued to depend on the excess charge available.[40] To be an effective tool in high-throughput proteomics, it is necessary both that the mass spectral signals of the individual proteins are high enough for detection and that the signal intensities for different protein types are as close as possible to each other for a given concentration.

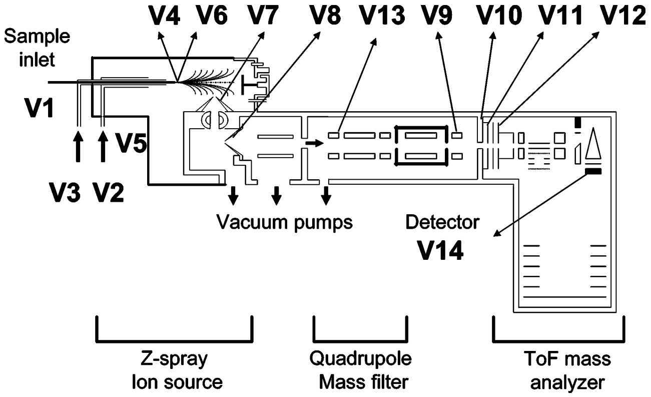

Figure 2. Schematic of the ESI-QTOF mass spectrometer used in this study, showing the Z-spray arrangement and the areas associated with the 14 variables (V1−V14, see Table 1 for details) optimized in the study.

Such a desired result may depend on many factors, such as solvent composition, instrumental parameters, and concentration effects, as it is known that these factors influence the sensitivity of detection and the charge-state distribution of a protein, even when analyzed in isolation.[10−13,15,16] We considered the optimization of relevant instrumental parameters that influence ionization and ion transmission to the analyzer, to see whether the signal intensities can be improved to provide an even signal output for the five proteins in a mixture. Table 1 indicates the 14 instrumental parameters that we chose to study and thus varied (such that we shall often refer to them as “variables” rather than parameters) and that may thus bear on this. Some of these adopt fixed (categorical) values, and some are continuous. A schematic of the mass spectrometer, showing the “locations” of the chosen variables is shown in Figure 2. If on average it is taken that they could each adopt 10 values, the search space is then 10^{14} experiments (and the lifetime of the Universe is $\sim 10^{17}$ s[41]).

We chose to effect this optimization by performing a genetic search. Figure 3A shows a plot of the response (fitness) distributions for the 14 variables, obtained from the experiment. It indicates the apparent multimodality of the fitness for some variables and the sharpness of the peaks in the fitness for others. We note in particular that the desolvation gas flow rate (V2), the source temperature (V4), and the transport hexapole voltage (V9) appear to have a multimodal distribution (although we recognize that the search space is sampled very sparsely), while that of the differential pumping aperture voltage (V10) optimizes to a unique value and that of the detector voltage (V14) optimizes to the maximum value available. Even in cases where a unique value for a variable is optimized, the distribution of the response ranges from the low to a high value, indicating the multivariate nature of the problem (i.e., the effect of a given variable depends significantly on the values of the other variables). To understand the nature or “ruggedness” of the search space, often referred to as

[40] Pan, P.; McLuckey, S. A. *Anal. Chem.* **2003**, *75*, 1491−1499.

[41] Barrow, J. D.; Silk, J. *The left hand of creation: the origin and evolution of the expanding universe*; Penguin: London, 1995.

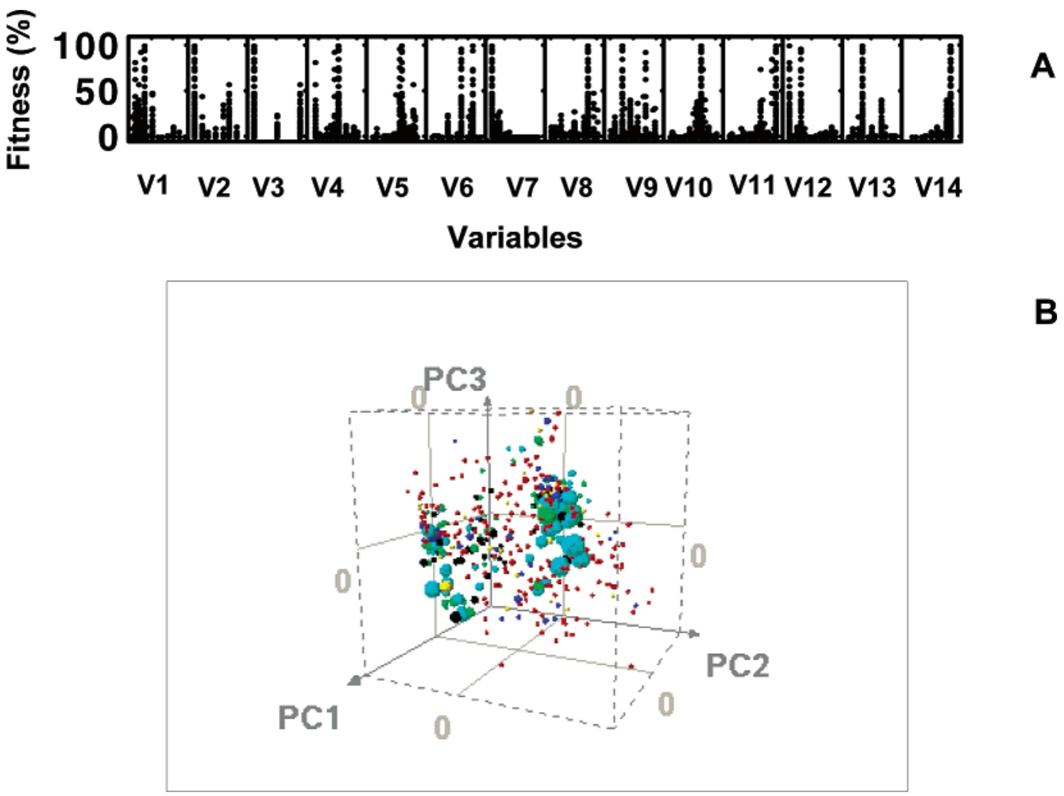

Figure 3. (A) Distribution of the relative fitness (%) for the 14 variables (V1−V14) (data obtained from the experimental results) and (B) principal component analysis clustering of the overall population with respect to generation and fitness function plotted as a pseudo 3D plot of the first three PCs extracted. The generation number is coded by color (1, red; 2, blue; 3, yellow; 4, black; 5, green; 6, cyan), and the fitness is encoded by the size of the individual points. Refer to Table 1 for the ranges of the variables.

the “fitness landscape”,[42−45] and to assess the progress of the GA with respect to this search space, we performed principal component analysis[46] on the range-scaled variables (all variables scaled to a range of 0−1), throughout the six generations. The first three principal components (explaining 52% of the variance) are plotted in Figure 3B, where the generation is encoded by color and the fitness by the size of the individual. Although we have populated the search space only very sparsely indeed (439 individuals in a nominal search space of 10^{14}), the implication is that there are a small number of very restricted areas in the search space that encode a good fitness and that the genetic search method is finding them.

Figure 4 shows how the fitness progressively improved both for the overall fitness per se (Figure 4A) and in terms of the contribution of the individual proteins to it (Figure 4B). Thus, the fitness increases significantly by generation 6, by which time the fitness of just 439 individuals had been evaluated. In particular, the median fitness per generation improved 400-fold, from 0.07% in generation 1 to 28% in generation 6, while the best individual improved some 5-fold (Figure 4A). Ideally, a “perfectly fit” individual should have a contribution of one-fifth from each protein, and it can be seen from Figure 4B that the fitter individuals are indeed approaching this. It can also be seen that the number of relevant peaks (those corresponding to the protein charge states) increases significantly as the fitness increases (Figure 4C).

We also used a genetic programming approach to evolve a function tree that best described the fitness. A tree that regularly evolved used the ratio or differences of the skimmer 1 (sample cone) (V7) and the skimmer 2 (extraction cone) (V8) voltages, and the effect of the latter on fitness is shown in Figure 4D. We can now understand that a substantial contribution to the fitness comes from holding the skimmer 1 voltage at a very low value, while maintaining the skimmer 2 voltage close to but not equal to this value. The optimal difference between V7 and V8 is typically just 2 V.

Figure 5 shows a comparison of the mass spectra of the protein mixture and the corresponding variable setting (normalized values) for the median individuals (i.e., the spectrum obtained by taking the median of all individual spectra for the given generation) in the first (Figure 5A) and sixth (Figure 5B) generations, and the best individual finally obtained (Figure 5C). Improvements in the spectral pattern with respect to the evenness of peak distribution can be clearly seen, by comparing the percentage of signal contributed by each of the five proteins, in the three spectra, shown as insets. The difference in the settings can also be noted. It must also be noted that although only the best individual is shown, there are more than one combination of settings that gave rise to a close to 100% fitness (i.e., cases where the proteins were more evenly detected). Finally, Figure 6 shows a distribution of the population (individuals from all six generations, normalized to the total and expressed as a percentage), with

[42] Kauffman, S. A. *The origins of order*; Oxford University Press: Oxford, 1993.
[43] Stadler, P. F. *J. Math. Chem.* **1996**, *20*, 1−45.
[44] Reeves, C. R. *Ann. Oper. Res.* **1999**, *86*, 473−490.
[45] Voigt, C. A.; Kauffman, S.; Wang, Z. G. In *Advances in Protein Chemistry*; Arnold, F. M., Ed.; Academic Press: San Diego, CA, 2001; Vol. 55, pp 79−160.
[46] Jolliffe, I. T. *Principal Component Analysis*; Springer-Verlag: New York, 1986.

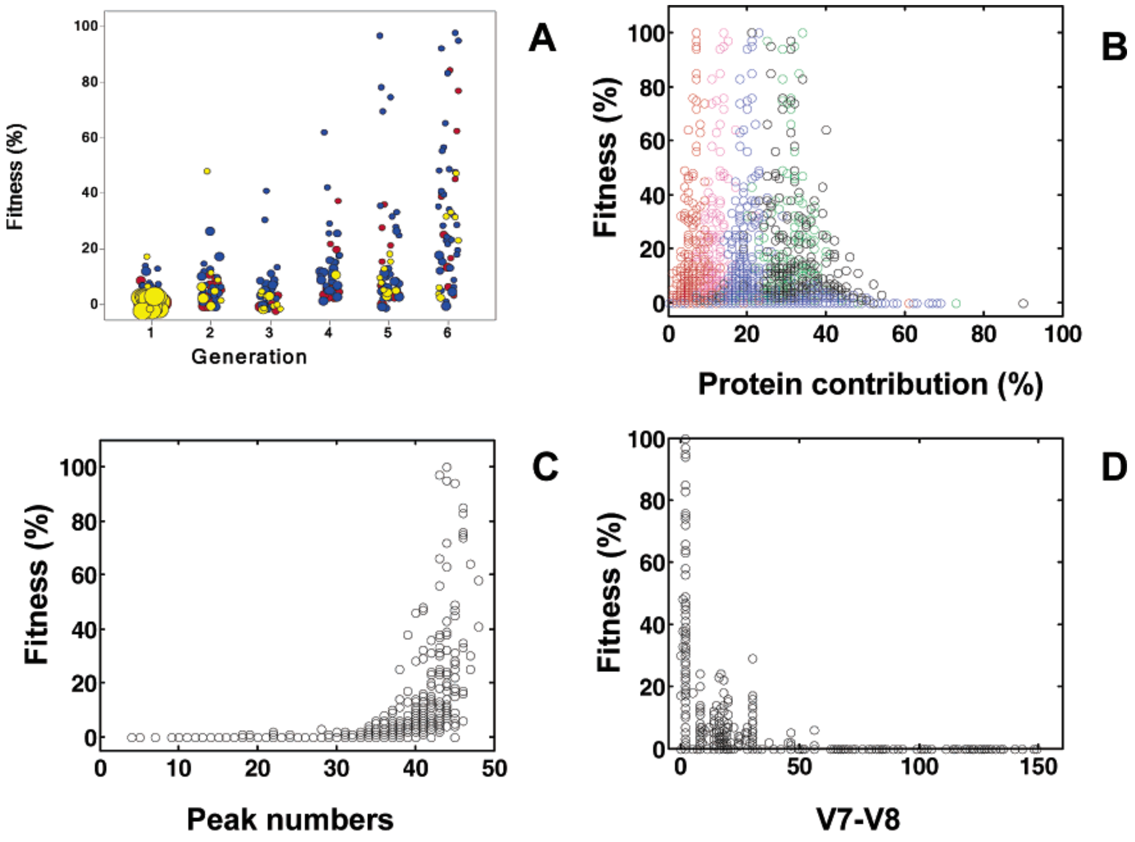

Figure 4. (A) Distribution of the relative fitness for each generation. Flow rate is coded by the color (20−60 μL/min, red; 60−200 μL/min, blue; 200−500 μL/min, yellow) and cone voltage (variable V7) coded by the size of the individual points, with increase in size indicating increase in cone voltage. (B) Distribution of the fitness with respect to the relative intensity contribution (I_y/I) for the individual proteins (insulin, magenta; ubiquitin, black; cytochrome *c*, green; lysozyme, red; myoglobin, blue). (C) Total peak numbers for the individuals of the population from the six generations with respect to the relative fitness. (D) Difference between the skimmer voltages (V7 − V8) for the individuals of the population from the six generations with respect to the relative fitness.

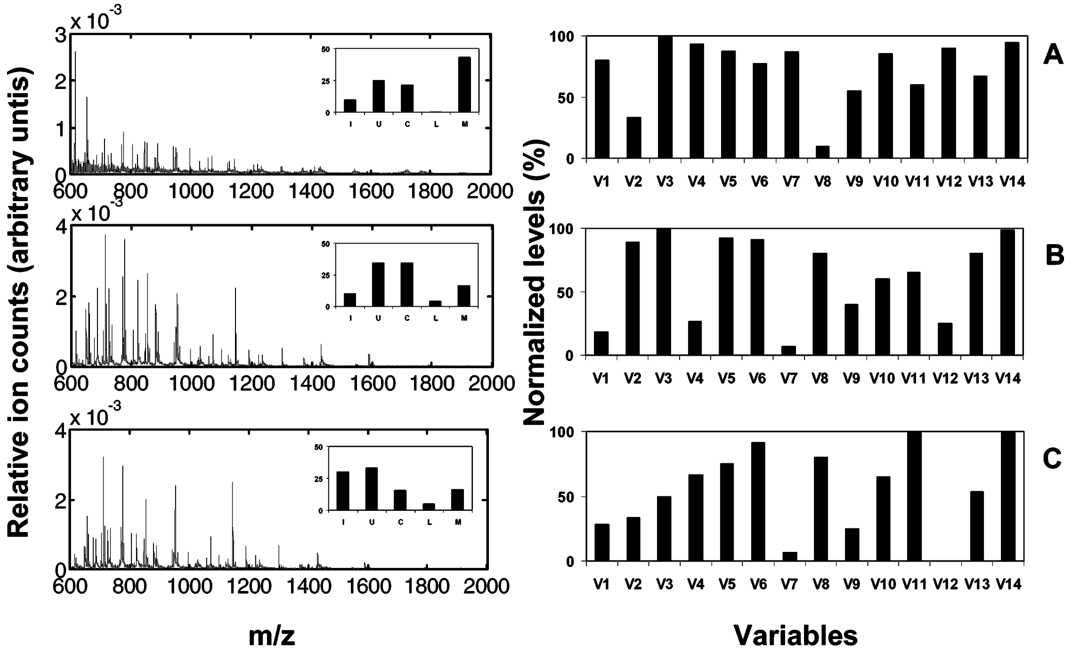

Figure 5. Positive ion ESI mass spectra of protein mixtures and the corresponding instrumental settings (variables) from individuals with a median fitness (i.e., spectrum obtained by taking the median of all individual spectra for the given generation) for the first (A) and the sixth (B) generations, and for the overall best individual (C), corresponding to a relative fitness of 0.07, 28, and 100%, respectively. The percentage of signal contributed by each of the five proteins (I , insulin; U, ubiquitin; C, cytochrome *c*; L, lysozyme; M, myoglobin) in the three spectra is shown as inset. The levels of the variables are normalized to the maximum value for each variable, for ease of visualization.

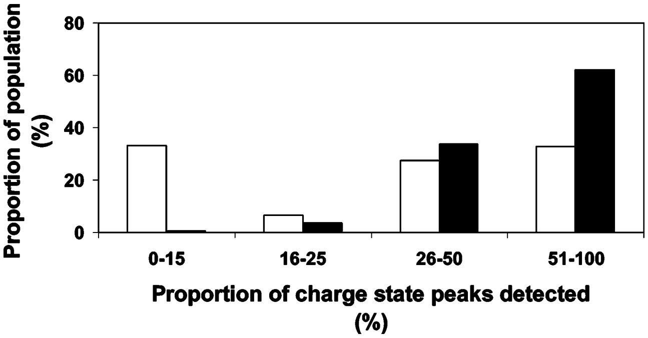

Figure 6. Distribution of the proportion of individuals for the first (white) and the sixth (black) generations, showing the number of charge-state peaks detected (as a percent of those expected relative to those seen in the "ideal" spectrum).

respect to the proportion of charge-state peaks detected (as a percentage of those expected in relation to the "ideal" spectrum) in generations 1 (white) and 6 (black). The proportion of charge-state peaks detected has improved significantly in generation 6, as a greater proportion of the individuals in this generation approach the desired 100% mark.

DISCUSSION AND CONCLUSIONS

Although they are not commonly posed as such, many experimental problems are in fact, or can best be cast as, combinatorial optimization problems. In these problems, there is a large search space where we may hope to find a good solution among what is often an astronomically large number of possible solutions. In the present work, we treated the optimization of instrumental conditions for the analysis of protein mixtures using ESI-MS as a combinatorial optimization problem and studied the application of a genetic search method in finding (an) optimal solution(s).

In the postgenomic era, there is great emphasis on studying cellular processes on a "global" scale to gain a systems-level understanding. As a consequence, analysis of intact protein mixtures using ESI-MS for high-throughput operations is driven by the need to obtain as complete a representation of proteomes as possible, with minimal recourse to cleanup and separation stages. In ESI-MS, the spectral signal(s) of a given protein depend(s) on the solvent the protein is dissolved in[12] and the ionization conditions employed and will differ for different instrumental conditions.[15,16] For proteins present in mixtures, matrix effects (due to the presence of other proteins or other ionizable species) might also influence the signal intensity.[40] Although not extensively (and certainly not systematically) investigated or understood, the instrumental conditions play an important role in influencing the ionization pattern and detectability of proteins in mixtures. Accordingly, we observed that a mixture of five proteins yielded a mass spectrum that was significantly different from the combined spectrum of the individual components (Figure 1) and that the ion intensity of the mixture spectrum was lower than that of the combined (theoretical mixture) spectrum under a given set of conditions. This can be attributed to the nature of the proteins and their gas-phase partitioning and ionization while being electrosprayed under the given set of conditions. Consequently, our investigation concentrated on the application of a genetic search algorithm toward finding optimal condition(s) in order to maximize the response from all the proteins and thus enhance detection of multiple protein species within a spectral acquisition. We chose five standard proteins in the mass range of 5−20 kDa. A major proportion of proteomes, at least for prokaryotes recorded in databases,[47] is composed of proteins in this mass range and has been shown to be amenable to MS analysis.[48] The proteins were dissolved in a solvent commonly employed for positive ion ESI-MS detection.[39]

The instrumental parameters we chose to optimize (Figure 2, Table 1) are those that would influence ionization (V1−V6), ion selection and transmission (V7−V13), and detection (V14). Some of these (and other) variables have been reported to influence the analysis of proteins by ESI-MS. For instance, raising the cone potential (V7 and V8) has been associated with shifts in the charge-state distributions of proteins.[14,15] In fact, collision-induced dissociation, achieved by varying the skimmer voltages, is a mode for generating controlled fragmentation of intact proteins to peptides in the "top-down" strategy for proteome characterizations.[8] It is also known that increasing metal capillary temperature improves the desolvation process, resulting in sharpened peaks.[13] Oberacher et al.[16] indicated that parameters at the ion source, such as the capillary voltage (V6), capillary temperature (V4), or sheath gas flow rates (V2,V3), influenced the protein mass spectral signals less than did settings for the ion-transfer optics (equivalent to V7−V13 here). Robinson and co-workers were able to observe intact ribosomes by employing a carefully balanced regime of pressure gradients throughout the mass spectrometer.[4] For the analysis of proteins in mixtures, such observations on individual variables may not always be generally applicable; since each protein may be differentially influenced by the different variables, the optimization should be addressed as a multivariate problem. This is evident from the way the fitness in our study varies with respect to each variable alone (Figure 3A) and together (Figure 3B).

The genetic search method adopted evolved parameters according to the chosen fitness function within six generations (Figure 4A), corresponding to an even detection of protein signal with respect to intensities (Figure 4B) and to peak numbers (Figure 4C), resulting in an improved detection of the different protein charge states (Figure 5A−C). It is also clear from Figure 6 that the proportion of the population showing a better detection of the charge-state peaks increases as we move from generation 1 (where a significant proportion (>30%) of the population had a low percent (0−15%) of charge-state peaks detected) to generation 6 (where a majority (>60%) of the population had a high percent (51−100%) of charge-state peaks detected). It must be noted that the variable levels of the individuals shown in Figure 5 are only representative. There were more than one individual close to the 100% fitness criterion, and these had some variable levels not very identical to the one shown for the best individual. Since the settings are dependent on each other, it is difficult to generalize on the optimized values.

However, following the optimization process, it was possible to comment on the variable values generally preferred when (non)-linear mapping between input and output variables was considered

[47] Demirev, P. A.; Ho, Y. P.; Ryzhov, V.; Fenselau, C. *Anal. Chem.* **1999**, *71*, 2732−2738.

[48] Johnson, J. R.; Meng, F. Y.; Forbes, A. J.; Cargile, B. J.; Kelleher, N. L. *Electrophoresis* **2002**, *23*, 3217−3223.

by genetic programming, as well as direct observation of the data. For instance, although lower cone voltages minimize fragmentation and maximize intact protein signal intensities, too low a voltage would influence the differential pumping of ions and hence the sensitivity of detection. However, in the present work, the genetic algorithm search was optimizing toward very low values of skimmer 1 (sample cone) voltage, values at which the individual proteins would be detected with relatively poor sensitivity, but above a threshold (10 V) preferred a high enough skimmer 2 (extraction cone) voltage that would minimize the potential difference between these two skimmers. It is possible that under these conditions ion transmission is less selective with respect to its influence on the individual proteins, allowing a majority of the five proteins to fly through and be detected. To our knowledge, this kind of a relationship between the skimmer voltages and the even detection of proteins has not been reported before. Although it might be argued that this relationship may be specific to the particular set of proteins investigated, it demonstrates nicely the hypothesis, or knowledge-generating ability of genetic search in this optimization process, under conditions where no a priori knowledge of the experimental conditions was used.

In summary, we noted that quite small changes in the settings on a mass spectrometer can have substantial effects on the effectiveness, and in particular the *differential* effectiveness, of the mass spectral response of proteins in mixtures. We treated the search for a good set of mass spectrometer settings as a combinatorial optimization problem and showed that a genetic search algorithm was extremely effective in finding instrumental conditions that resulted in significant improvements in the response over the starting set. Although the search space was a nominal 10^{14} experiments, the genetic search was found to converge rapidly toward optimal solutions even when only 439 of them had been evaluated. Such inductive, evolutionary, machine learning approaches to experimental design, and optimization might in the future beneficially be automated in a closed loop form, as has been done in fields such as laser spectroscopy [49−51] and functional genomics.[52]

ACKNOWLEDGMENT

We thank the U.K. BBSRC and EPSRC for financial support.

Received for review June 20, 2003. Accepted September 12, 2003.

AC034669A

[49] Judson, R. S.; Rabitz, H. *Phys. Rev. Lett.* **1992**, *68*, 1500−1503.

[50] Levis, R. J.; Menkir, G. M.; Rabitz, H. *Science* **2001**, *292*, 709−713.

[51] Daniel, C.; Full, J.; Gonzalez, L.; Lupulescu, C.; Manz, J.; Merli, A.; Vajda, S.; Woste, L. *Science* **2003**, *299*, 536−539.

[52] Bryant, C. H.; Muggleton, S. H.; Oliver, S. G.; Kell, D. B.; Reiser, P.; King, R. D. *Electron. Trans. Artif. Intelligence* **2001**, *5*, 1−36 (http://www.ep.liu.se/ej/etai/2001/2001/).