

## Strategy to Design Improved Proteomic Experiments Based on Statistical Analyses of the Chemical Properties of Identified Peptides

Martin Ethier and Daniel Figeys\*

*Ottawa Institute of Systems Biology, Ottawa, Canada*

Received August 30, 2005

Proteomics is an emerging field that uses many types of proteomic platforms however has few standardized procedures. Deciding which platform to use to perform large-scale proteomic studies is either based on personal preference or on so-called "figures of merit" such as dynamic range, resolution, and the limit of detection; these factors are often insufficient to predict the outcome of the experiment as the detection of peptides correlates to the chemical properties of each peptide. There is a need for a novel figure of merit that describes the overall performance of a platform based on measured output, which in proteomics is often a list of identified peptides. We report the development of such a figure of merit based on a predictive genetic algorithm. This algorithm takes into account the properties of the observed peptides such as length, hydrophobicity, and *pI*. Several large-scale studies that differed in sample type or platform were used to demonstrate the usefulness of the algorithm for improved experimental design. The figures that were obtained were clustered to find platforms that were biased in similar ways. Even though some platforms are different, they lead to the identification of similar peptide types and are thus redundant. The algorithm can thus be used as an exploratory tool to suggest a minimal number of complementary experiments in order to maximize experimental efficiency.

**Keywords:** mass spectrometry • proteomics • figures of merit • experimental design • genetic algorithm

### Introduction

Various combinations of instruments and experimental procedures are available for large-scale proteomics studies.<sup>1–7</sup> To obtain meaningful results, a lot of time needs to be spent on experimental design, i.e., finding an optimal platform to use to maximize the number of identified proteins contained in a complex sample. In many cases, researchers base their choice on classical analytical figures of merit such as the limit of detection, dynamic range and sensitivity<sup>8–11</sup> or even the number of proteins identified.<sup>12,13</sup> While these may provide the best overall analytical signal, it does not take into account the experimental results and the fact that some part of any given experimental procedure prior to mass spectrometric detection might bias the specific types of peptides available for detection. A more meaningful way of choosing an analytical technique would be to consider the peptides that are positively identified and try to deduce a combination of platforms which would provide the maximum coverage of peptide types. For the purposes of this study, a proteomic platform is considered to be a "black box" where a biological sample enters and peptide identifications exit. Within a given platform, an analyte may be manipulated in numerous ways including protein depletion,

protein modification, protein separation, and peptide separation; the peptides are commonly detected using mass spectrometry.

To improve overall analytical efficiency, it would be beneficial for experimentalists to define a set of available platforms and run a standardized sample on each of them. The peptides identified from these experiments would then be analyzed and figures of merit would be assigned to each platform using the proposed algorithm in this study, which extracts global figures of merit, representative of the trends observed in the chemical properties (such as length, *pI* and hydrophobicity) of the peptides identified. Any similarities between the distributions of these properties would suggest that these techniques are most suitable to the study of these types of peptides. The user would then have more information to base their choice of which platform(s) to use.

The developed algorithm based on the widely used genetic algorithms principle<sup>14</sup> and cluster analysis<sup>15</sup> is called the Predictive Genetic Algorithm (PGA). The Plasma Proteome Project (PPP)<sup>16</sup> dataset made freely available by the Human Proteome Organization (HUPO) was used as a dataset to simulate a set of available platforms. The figures of merit derived in this study were then used to classify the different proteomic platforms used in the PPP based on their peptide outputs and make an intelligent prediction of which platform(s) would be most suitable for a similar type of analysis in the future. Other datasets freely available in the literature were also

\* To whom correspondence should be addressed. Ottawa Institute of Systems Biology, University of Ottawa, 451 Smyth Road, Ottawa, Ontario, Canada K1H 8M5. Phone: 613-562-5800 ext 8674. Fax: 613-562-5452. E-mail: dfigeys@uottawa.ca; www.oisb.ca.

**Table 1.** Description of Proteomic Platform Studied as Part of the PPP<sup>a</sup>

lab ID	depletion	protein separation	reduction/alkylation	peptide separation	mass spectrometry	search software
1	aig	none	iam	rp/scx/rp	ESI-MS/MS (decasp)	PEPMINER
2	none	cho affinity	iam	scx/rp	ESI-MS/MS (qtof)	SEQUEST
11	none	cho affinity	iam	rp	ESI-MS/MS (decasp)	SEQUEST
12	aig	none	iam	rp/scx/rp	ESI-MS/MS (deca)	SEQUEST
17	aig	1d sds	iam	rp	ESI-MS/MS (lcq)	SEQUEST
21	top6	rotofor-ief/rp/1d-sds	iam	rp	ESI-MS/MS (qtof)	MASCOT
	top6	rotofor-ief/rp/1d-sds	none	none	MALDI-MS/MS (abi4700)	MASCOT
	top6	rotofor-ief/rp/1d-sds	none	rp	ESI-MS/MS (qtof)	MASCOT
22	top6	1d sds	iam	rp/scx/rp	ESI-MS/MS (decasp)	SEQUEST
28	ig	none	none	rp	ESI-FTICR	VIPER
29	top6	none	iam	scx/rp	ESI-MS/MS (decasp)	SEQUEST
	top6	none	iam	scx/rp/2mz	ESI-MS/MS (decasp)	SEQUEST
34	top6	zoom-ief/1d-sds	iam	rp	ESI-MS/MS (decasp or ltq)	SEQUEST
40	none	aig affinity/rp	iam	scx/rp	ESI-MS/MS (lcq)	SONAR
43	aig	none	iam	rp	ESI-MS/MS (qtof)	MASCOT
	aig	none	iam	rp	MALDI-MS/MS (abi4700)	MASCOT
55	none	sax	iam	rp	ESI-MS/MS (ltq)	SEQUEST

<sup>a</sup> Only platform having identified more than 250 unique peptides were included. Abundant protein depletion: a = albumin, ig = immunoglobulin and top6 = agilent depletion product. Protein Separation: cho = carbohydrate, ief = isoelectric point focusing, rp = reverse phase and sax = strong anion exchange, sds = sodium dodecyl sulfate. Alkylation: iam = iodoacetamide. Peptide Separation: scx = strong cation exchange.

**Table 2.** Description of Experiments Publicly Available in the Literature<sup>a</sup>

platform	sample origin	protein separation	reduction/alkylation	peptide separation	mass spectrometry	search software
Le Bihan	breast cancer total cell lysis and membrane	none	iam	rp	ESI-MS/MS (QStar)	MASCOT
Kristensen	breast cancer membrane	none	iam	rp	ESI-MS/MS (QStar)	MASCOT
Chan	serum	none	none	ief/scx/rp	ESI-MS/MS (decasp)	SEQUEST

<sup>a</sup> iam = iodoacetamide, rp = reverse phase, ief = isoelectric focusing, scx = strong cation exchange.

incorporated in this work to demonstrate additional capabilities and extensions of this algorithm.

## Experimental Section

**Plasma Proteome Project.** The list of peptides identified used as the training set were obtained from the publicly available HUPO dataset for the plasma proteome project (PPP). The sample was prepared by HUPO<sup>16</sup> in the following way. Blood from pooled male/female samples was divided into serum (b1-serum) and plasma. The plasma was further treated with anticoagulants: Citrate (b1-cit), heparin (b1-hep), and EDTA (b1-edta). An aliquot of the sample was provided to any lab wanting to participate in the project. All information available for each proteomic platform used by the participating labs is shown in Table 1. No specific information such as gradient length or mass spectrometric acquisition cycle was provided. For proper sampling, only platforms reporting more than 250 unique peptides were used in the present study.

**Other Large-Scale Studies.** Datasets from other human sources were found in the literature and also used to further test the algorithm.<sup>17–19</sup> The description of these studies is shown in Table 2.

**Theoretical Trypsinome.** The nonredundant human protein database (PIR-NREF version 1.53) was downloaded from the Protein Information Resource (PIR) web site.<sup>20</sup> The theoretical tryptic digestion was performed on all 28 625 proteins using Proteogest.<sup>21</sup> Peptide redundancy was removed using software written in-house to yield over 600 000 unique peptides.

**Bioinformatics.** The software suite for the statistical analysis, the predictive genetic algorithm and the clustering was developed in-house using C++Builder 5 (Borland, Scotts Valley, CA).

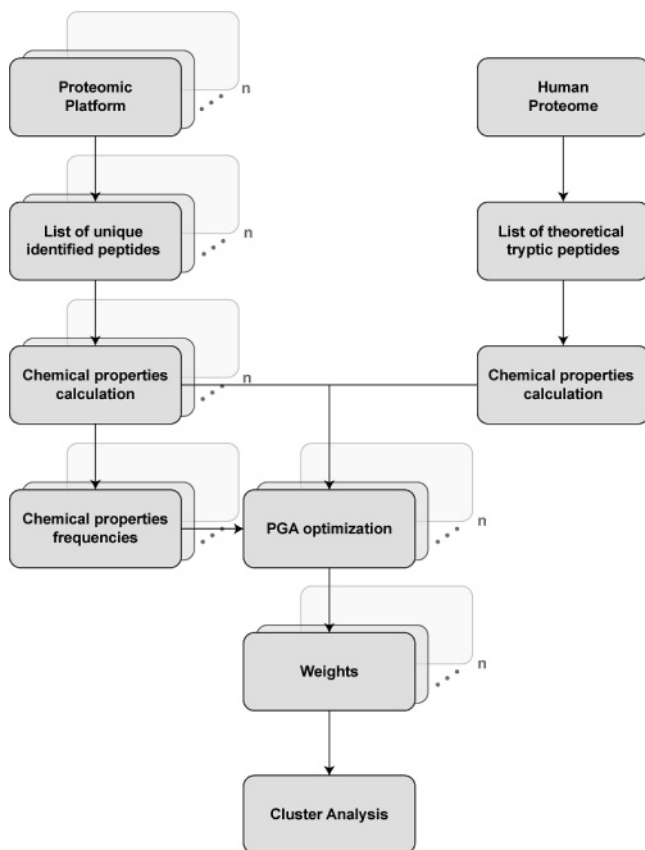
The application and example files are available online at [www.oisb.ca](http://www.oisb.ca).

## Results and Discussion

**Predictive Genetic Algorithm.** The general approach used by the algorithm for the treatment of any proteomic platform as part of an experimental design study is shown in Figure 1. For each observed peptide, the values of some of its chemical properties are calculated: length, *pI* (as predicted by the *pK<sub>a</sub>* values of the free amino acids<sup>22</sup>) and hydrophobicity (as predicted using the amino acid weights calculated by Le Bihan et al.<sup>19</sup>). The same is done to the theoretical list of unique peptides from the human trypsinome. These properties were chosen because of their ease of prediction. Length was chosen over *m/z* because it is more meaningful to peptide separation without losing the relation to mass spectrometric detection as length is related to molecular weight as shown in Figure 2.

The hypothesis behind the algorithm is that for a given proteomic platform, if there is a bias toward a chemical property, there should be a difference in the frequency distribution of that property for that proteomic platform compared to its frequency distribution in the theoretical human trypsinome. To verify this hypothesis, the frequency distribution of each chemical property is calculated for the platform and the theoretical trypsinome. To reflect the importance of each property in differentiating peptides for that particular proteomic platform, weights are calculated. A score distribution is calculated using both these weights and the frequency distributions as shown in eq 1.

$$\text{Score}_{\text{pept}} = W_L \cdot \text{Freq}_L^{1/3} + W_H \cdot \text{Freq}_H^{1/3} + W_{pI} \cdot \text{Freq}_{pI}^{1/3} \quad (1)$$



**Figure 1.** Schematic representation of the predictive genetic algorithm. Each of the  $n$  platform is analyzed independently until the cluster analysis is performed on all weights obtained. Unique peptides identified are first extracted. Length, hydrophobicity and  $pI$  of each peptide are calculated. This process is also done once for the human trypsinosome. The frequency distribution of each property is calculated for each platform. Weights are determined for each platform based on the comparison of the platform and the human trypsinosome results using the predictive genetic algorithm.

where  $W_x$  and  $Freq_x$  are the weight and frequency distribution assigned to property  $X$  for a given experiment, respectively.

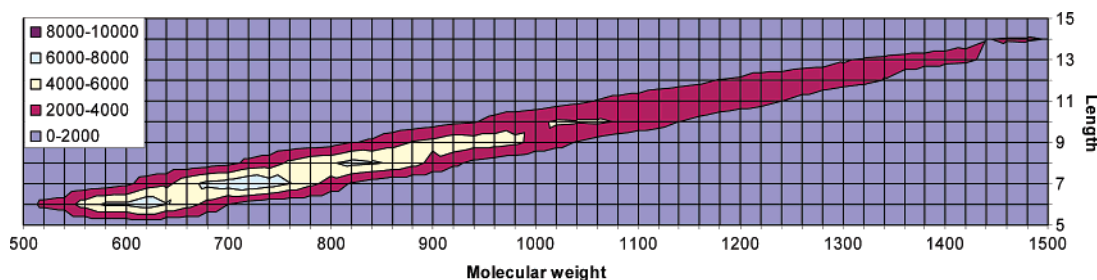
This equation was modified from the equation presented by Le Bihan et al.<sup>19</sup> In their study, Le Bihan et al. were multiplying all terms together, which puts equal weight on each property losing the information about any property bias. In this study, the terms are added together using different weights representing the bias observed. However, the appropriate weights to use are not known a priori, they need to be determined through optimization calculations. This was accomplished by finding the weights that led to the maximum root-mean-square dif-

ference between the score distribution for the platform and the theoretical trypsinosome; a genetic algorithm was used for this purpose.<sup>14</sup>

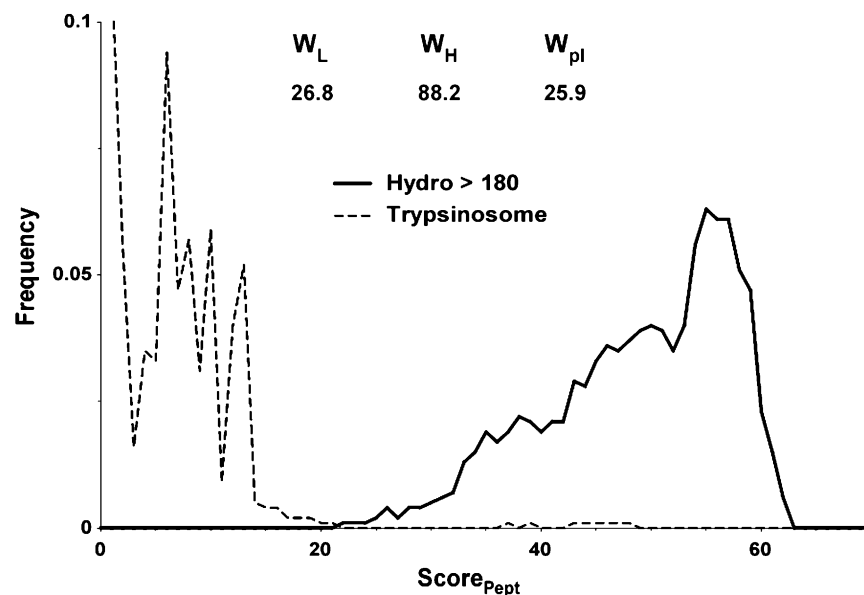
In genetic algorithms, the naming convention and definition of elements was historically based on genetics because of similarities in the principles of the algorithm and natural selection in genetics. For example, an iteration (or cycle) of the genetic algorithm is called a "generation". A set of weights to be tested is called a "gene". All genes in a given iteration are called a "gene pool". In each generation, the "quality" of each gene in the gene pool is investigated, which means that the difference between the score distributions of the platform and the trypsinosome is calculated for each set of weights for a given iteration. The first generation of genes is randomly generated. The quality of each gene is evaluated and sorted accordingly. Only the two "fittest" genes of the gene pool are able to survive and used to "breed" the next generation and are used to generate the new gene pool, which means that the two sets of weights leading to the maximum difference are used in the next iteration and linear combinations of these sets are used to generate the other sets of weights in that iteration.

In the present study, an optimization is composed of 20 generations. Each generation has a gene pool of 12 genes. Each gene pool is composed of three parent genes (two fittest genes from the previous generation and one random gene), three "offspring" genes generated from a combination of the parents, three "mutants" genes generated from one parent and three random genes. To ensure that the optimization converges quickly and consistently toward the global maximum, three independent optimizations with different random starting gene pools are performed. The four fittest genes of each optimization are combined together to be the new starting gene pool of a fourth and last optimization. The resulting fittest gene is considered to be the optimized set of weights.

The importance of each property in differentiating the distributions is thus reflected by the weight obtained after the optimization. By comparing the weights, it is possible to evaluate which property was biased for that proteomic platform. To test the ability of the algorithm to extract meaningful weights, a subset of the trypsinosome was analyzed using the algorithm. Peptides having a hydrophobicity larger than 180 were extracted from the trypsinosome without any selection of  $pI$  and length. This led to a subset of 995 peptides. The subset was analyzed using the algorithm and the resulting frequency distributions for the peptide scores of the platform and the theoretical trypsinosome are shown in Figure 3. It can be observed that the weight for hydrophobicity is almost four times higher than the other properties. The frequency distribution for this subset was easily differentiated from the whole trypsinosome, which suggests an important observed bias.



**Figure 2.** Distribution of the peptide length and molecular weight for the theoretical trypsinosome showing the linear relationship.

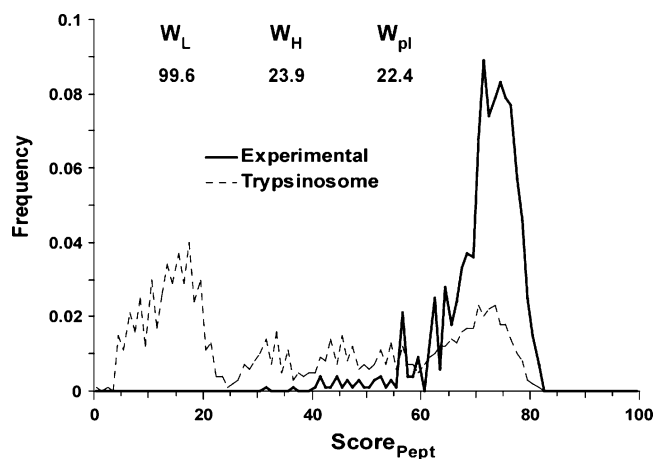


**Figure 3.** Frequency distribution of the peptide scores calculated using eq 1 and the weights determined using the predictive genetic algorithm on a subset of peptides of the trypsinosome having an hydrophobicity larger than 180. The human trypsinosome peptide score frequency distribution is also shown. The weights obtained are the one leading to the maximum root-mean-square difference between these two distributions.

**Cluster Analysis.** Once the importance of each chemical property is determined for each platform, experiments leading to similar biases can be regrouped using a cluster analysis algorithm<sup>15</sup> based on the observed weights. The clustering is done using the euclidian distance, consisting of the root-mean-square difference of weights, to measure the difference between sets of weights. Clusters are generated using the unweighted pair-group centroid and a threshold of 60 (empirically determined based on the PPP results) to classify significantly different groups. At each iteration of the clustering algorithm, the two closest clusters are grouped together to form a new cluster for the next iteration. The un-weighted pair-group centroid represents an average of each corresponding property for two sets.

**Classification of Platforms: HUPO Dataset.** The PPP dataset was created from an initiative of the HUPO organization. The goals of the study were to provide scientists with an exhaustive list of proteins in human serum and plasma and observe any variations due to subject characteristics (age, sex, etc...). This dataset was used as a training set for our genetic algorithm. Because the plasma proteome project analyzed the same sample using different techniques the difference in peptides identified will depend on the analytical technique and experimental procedures used and not the sample origin. This dataset can thus be used to simulate an experimental design study where each platform from the PPP is considered an available platform for our virtual lab. It is assumed that each platform is optimized and stable. The results obtained are a snapshot at the time of analysis.

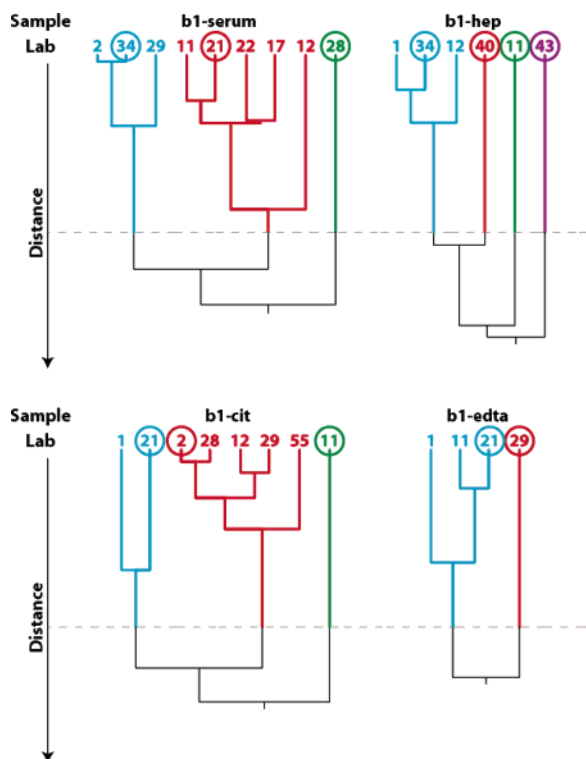
The weights for each proteomic platform were determined using the predictive genetic algorithm. Figure 4 shows an example of the results of a typical PGA optimization where a large difference is observed between the score distributions of the platform and the trypsinosome. The weights calculated are the ones leading to the largest differentiation between experimental peptides and the theoretical peptides. It may be observed that the majority of the peptides in the experimental



**Figure 4.** Peptide score frequency distribution calculated after using the predictive genetic algorithm on platform 29 analyzing the b1-serum sample. If it is possible to find a set of weights bringing an important difference, it means a bias of peptide types is observed.

datasets have high scores, whereas the theoretical peptide scores have a wider distribution and a smaller fraction of them have high scores. Thus, the technique efficiently separates the peptides that were shown experimentally to have a biased probability of being observed. The weights can then give an indication of the importance of the bias of each chemical property for this particular proteomic platform.

The resulting weights were clustered to find experimental bias toward the same type of peptides. Figure 5 shows the results of the cluster analysis for each sample. Some platforms (for example see Labs 1 and 34), while being quite different are biased toward the same peptide types. Looking at the experimental description alone, this could not have been predicted. The number of experiments to perform was assessed by the relative distance between clusters. Table 3 shows the difference in chemical properties between the clusters.



**Figure 5.** Cluster analysis of platforms used in four different samples as analyzed using the PGA. All lab part of the same cluster are considered equivalent. The circled lab in each cluster is the lab leading to the most peptide identification. The dotted line represents the Euclidian distance of 60 used to classify significantly different clusters of platform.

**Table 3.** Clusters Observed at an Euclidian Distance of 60 for Each Sample<sup>a</sup>

Sample	Cluster Centroid			Best lab (peptides)
	Length	Hydro	pI	
b1-cit	89.1	81.9	12.4	21 (698)
	84.4	6.4	6.7	2 (1372)
	56.3	50.9	90.9	11 (300)
b1-edta	87.9	86.8	33.0	21 (694)
	95.6	10.1	15.4	29 (1195)
b1-hep	97.5	68.9	9.3	34 (1013)
	100.0	3.0	2.6	40 (1380)
	59.2	100.0	61.1	11 (338)
b1-serum	75.6	14.2	100.0	43 (341)
	99.3	14.4	13.1	34 (4473)
	75.4	81.6	34.3	21 (694)
b1-serum	98.5	6.9	98.8	28 (860)

<sup>a</sup> Each weight is the average of platforms part of that cluster. Best lab refers to the lab leading to the largest number of peptides identified in that cluster.

These results can be used to detect any similarities between seemingly different platforms when looking only at the procedure involved. A minimal number of platforms for maximum peptide coverage can then be defined. For every new platform, the sample is analyzed and the results compared to existing characterized platforms to see if it is significantly different to others already available. In this way, the analytical value of each

**Table 4.** Weights for the Chemical Properties as Calculated Using the Predictive Genetic Algorithm for Other Datasets Found in the Literature<sup>a</sup>

platform	cluster centroid			no.. peptides
	length	hydro	pI	
Le Bihan	100.0	12.2	9.3	1074
Kristensen	68.6	42.6	36.7	2461
Chan	100.0	4.0	3.0	2584

<sup>a</sup> Le Bihan and Kristensen used the same platform but a different sample. Chan studied a sample from human serum.

platform may be ascertained thereby reducing or eliminating redundant experimentation.

Some datasets extracted from the literature were studied using the algorithm. The weights that were determined are shown in Table 4. Chan et al.<sup>17</sup> also analyzed human serum and their data was used to simulate the addition of a new platform to our existing platform analysis. Following our analysis, their data seemed to resemble the first group of b1-serum in Table 3. One then has to decide if it is worth using a platform that duplicates the type of peptides that can be observed or even if it could replace the one currently used as the best representative of that group. This choice could then be based on other considerations such as ease of use, price or the number of peptides identified. In this particular case, our analysis suggests that the existing platform (34) is already optimal for maximum peptide coverage.

The identification of peptides is heavily influenced by their length, most likely due to limitations in the *m/z* range of mass spectrometers. Peptides outside these ranges are unlikely to be detected unless they are multiply charged as is the case in electrospray ionization. Furthermore, it has been demonstrated that doubly charged peptides commonly have fragmentation patterns that lead to stronger (and hence a larger number of) peptide identifications.<sup>23</sup> The ability of a peptide to be doubly charged will affect the length distribution by shifting it toward larger peptides. This should not be observed in the case of MALDI ionization, as MALDI generally produces singly charged ions;<sup>23</sup> however, this technique was under-represented in the PPP and never analyzed on its own. Therefore, we were not able to successfully evaluate the influence of the length bias for peptides analyzed with MALDI. On the other hand, the importance of hydrophobicity and pI varies greatly between platforms. These variations are probably due to the preparation and separation techniques used which varies greatly from platform to platform.

**Differentiation of Biological Samples.** The algorithm may also be used for a different type of experimental design study whereby different samples may be analyzed using one platform rather than varying the platform to use on one sample. This method is useful to characterize any part of a proteomic platform. One only needs to perform an analysis on two platforms differing by a particular part of the procedure. To illustrate this concept, two studies from the literature were analyzed.<sup>18,19</sup> In this example, both platforms are considered to be the same even though there are obviously some minor variations. In the study of Kristensen et al.<sup>18</sup> and Le Bihan et al.,<sup>19</sup> the blackbox ideology was thus extended to include the sample origin and treatment. This was done to study the effect of the different cell lines on the type of peptides observed but also the effect of membrane extraction. The weights determined from the Le Bihan et al. study show a predominance of length.

Furthermore, the weights of the Kristensen et al. study show an increase of the hydrophobicity and *pI* importance. This is most likely due to the fact that membrane proteins are more hydrophobic and few cytosolic proteins are present. The peptide sample in Le Bihan et al. was derived from a total cell lysate and must have had a more normal distribution of peptide hydrophobicity, as observed for human trypsinosome.

### Conclusion

The predictive genetic algorithm was applied to a large-scale study of the human plasma proteome to simulate an experimental design study. The results from a total of 13 different proteomic platforms were compared using a different figure of merit than the classical analytical figures of merit such as resolution, sensitivity and limit of detection. The resulting cluster analysis grouped the platform into different types. It was shown that several platforms overlapped in terms of the type of peptides they were able to identify. Without the PGA, these trends would be difficult to observe based on a description of the experiment alone. The algorithm was used to suggest a minimal number of platforms to use for maximum coverage of peptide types. This study shows that a lab could characterize all its available platforms and choose an optimal set of platforms and even determine if newly available platforms should be included in an analysis or even replace one or more of platforms available at the time.

The algorithm can also be used to directly observe the effect of any one experimental procedure in a platform by varying it within the platform and analyzing a single sample. This was done to study of the effect of membrane extraction on the nature of peptides observed and demonstrates the flexibility of the algorithm.

It was also observed that length is the most important property that affects peptide detection, as the weight assigned to length is always high. Due to the fact that hydrophobicity and *pI* vary to a larger extent from platform to platform, they are more useful properties for the differentiation of platforms.

**Acknowledgment.** The authors would like to thank Dr. Gilbert S. Omenn for giving us access to the HUPO dataset, Fred Elisma for making the PGA available online, and Dr. Jeffrey C. Smith for a critical review of the manuscript. This project was funded by MDS Inc., Protana Inc. and Genome Canada through the Ontario Genomics Institute.

### References

- (1) Aebersold, R.; Mann, M. *Nature* **2003**, *422*, 198–207.
- (2) Shevchenko, A.; Jensen, O. N.; Podtelejnikov, A. V.; Sagliocco, F.; Wilm, M.; Vorm, O.; Mortensen, P.; Shevchenko, A.; Boucherie, H.; Mann, M. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 14440–14445.
- (3) Opiteck, G. J.; Jorgenson, J. W. *Anal. Chem.* **1997**, *69*, 2283–2291.
- (4) Washburn, M. P.; Wolters, D.; Yates, J. R., 3rd *Nat. Biotechnol.* **2001**, *19*, 242–247.
- (5) Koller, A.; Washburn, M. P.; Lange, B. M.; Andon, N. L.; Deciu, C.; Haynes, P. A.; Hays, L.; Schieltz, D.; Ulaszek, R.; Wei, J.; Wolters, D.; Yates, J. R., 3rd *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 11969–11974.
- (6) Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. *J. Proteome Res.* **2003**, *2*, 43–50.
- (7) Ihling, C.; Sinz, A. *Proteomics* **2005**, *5*, 2029–2042.
- (8) Bjorhall, K.; Miliotis, T.; Davidsson, P. *Proteomics* **2005**, *5*, 307–317.
- (9) Wu, S. L.; Choudhary, G.; Ramstrom, M.; Bergquist, J.; Hancock, W. S. *J. Proteome Res.* **2003**, *2*, 383–393.
- (10) Lopez, M. F.; Berggren, K.; Chernokalskaya, E.; Lazarev, A.; Robinson, M.; Patton, W. F. *Electrophoresis* **2000**, *21*, 3673–3683.
- (11) Neuhoff, N.; Kaiser, T.; Wittke, S.; Krebs, R.; Pitt, A.; Burchard, A.; Sundmacher, A.; Schlegelberger, B.; Kolch, W.; Mischak, H. *Rapid Commun. Mass Spectrom.* **2004**, *18*, 149–156.
- (12) Gan, C. S.; Reardon, K. F.; Wright, P. C. *Proteomics* **2005**, *5*, 2468–2478.
- (13) Ostrowski, L. E.; Blackburn, K.; Radde, K. M.; Moyer, M. B.; Schlatter, D. M.; Moseley, A.; Boucher, R. C. *Mol. Cell Proteomics* **2002**, *1*, 451–465.
- (14) Goldberg, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*; Addison-Wesley: New York, 1989.
- (15) Sneath, P. H. A.; Sokal, R. R. *Numerical Taxonomy*; Freeman: San Francisco, 1973.
- (16) Omenn, G. S. *Proteomics* **2004**, *4*, 1235–1240.
- (17) Chan, K. C.; Lucas, D. A.; Hise, D.; Schaefer, C. F.; Xiao, Z.; Janini, G. M.; Buetow, K. H.; Isaaq, H. K.; Veenstra, T. D.; Conrads, T. P. *Clin. Proteomics* **2004**, *1*, 101.
- (18) Kristensen, D. B.; Brond, J. C.; Nielsen, P. A.; Andersen, J. R.; Sorensen, O. T.; Jorgensen, V.; Budin, K.; Matthiesen, J.; Venø, P.; Jespersen, H. M.; Ahrens, C. H.; Schandorff, S.; Ruhoff, P. T.; Wisniewski, J. R.; Bennett, K. L.; Podtelejnikov, A. V. *Mol. Cell. Proteomics* **2004**, *3*, 1023–1038.
- (19) Le Bihan, T.; Robinson, M. D.; Stewart, II.; Figeys, D. *J. Proteome Res.* **2004**, *3*, 1138–1148.
- (20) <http://pir.georgetown.edu/cgi-bin/nfspecies.pl?taxon=9606>.
- (21) [www.utoronto.ca/emililab/program/proteogest.htm](http://www.utoronto.ca/emililab/program/proteogest.htm).
- (22) Bjellqvist, B.; Hughes, G. J.; Pasquali, C.; Paquet, N.; Ravier, F.; Sanchez, J. C.; Frutiger, S.; Hochstrasser, D. *Electrophoresis* **1993**, *14*, 1023–1031.
- (23) Jonsson, A. P. *Cell Mol. Life Sci.* **2001**, *58*, 868–884.

PR0502900