Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Identifying differences in protein expression levels by spectral counting and feature selection

P.C. Carvalho[1], J. Hewel[2], V.C. Barbosa[1] and J.R. Yates III[2]

[1]Programa de Engenharia de Sistemas e Computação, COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brasil
[2]Department of Cell Biology, The Scripps Research Institute, La Jolla, CA, USA

Corresponding author: P.C. Carvalho
E-mail: carvalhopc@cos.ufrj.br

Genet. Mol. Res. 7 (2): 342-356 (2008)
Received January 18, 2008
Accepted February 7, 2008
Published April 15, 2008

ABSTRACT. Spectral counting is a strategy to quantify relative protein concentrations in pre-digested protein mixtures analyzed by liquid chromatography online with tandem mass spectrometry. In the present study, we used combinations of normalization and statistical (feature selection) methods on spectral counting data to verify whether we could pinpoint which and how many proteins were differentially expressed when comparing complex protein mixtures. These combinations were evaluated on real, but controlled, experiments (yeast lysates were spiked with protein markers at different concentrations to simulate differences), which were therefore verifiable. The following normalization methods were applied: total signal, Z-normalization, hybrid normalization, and log preprocessing. The feature selection methods were: the Golub index, the Student *t*-test, a strategy based on the weighting used in a forward-support vector machine (SVM-F) model, and SVM recursive feature elimi-

nation. The results showed that Z-normalization combined with SVM-F correctly identified which and how many protein markers were added to the yeast lysates for all different concentrations. The software we used is available at http://pcarvalho.com/patternlab.

Key words: MudPIT; Feature selection; Support vector machine; Spectral counting; Feature ranking

INTRODUCTION

A goal of proteomics is to distinguish between various states of a system to identify protein expression differences (Jessani et al., 2005). The first strategies used two-dimensional gel electrophoresis for comparing the migration of proteins according to their molecular weight and isoelectric point. In 2002, alternative approaches emerged to compare biological samples from different states. Mass spectrometry (MS) analysis was performed on enriched proteins that were fractionated on the surface of an MS plate. By correlating the peptide mass to charge (m/z) values obtained from SELDI-TOF MS (surface enhanced laser desorption ionization time-of-flight MS) with peptide abundance, Petricoin et al. (2002) used machine learning over a SELDI-TOF dataset acquired from SELDI-TOF MS of serum from control subjects and ovarian cancer patients. As a second step, unknown spectra were classified as belonging to the patient or control subject class (Petricoin et al., 2002; Unlu et al., 1997). A variety of feature selection/classification methods have since then been described as being used for this purpose, including genetic algorithms (Shah and Kusiak, 2004), Fisher criterion scores (Kolakowska and Malina, 2005), beam search (Badr and Oommen, 2006; Carlson et al., 2006), branch-and-bound (Polisetty et al., 2006), Pearson correlation coefficients (Mattie et al., 2006), and support vector machines recursive feature elimination (SVM-RFE; Carvalho et al., 2007).

The need for high sensitivity when analyzing samples of greater complexity led to the use of liquid chromatography (LC) coupled with electrospray MS (LC-MS) to profile digested protein mixtures. Elimination of the data-dependent tandem MS process enhances the detection of ions, since the instrument spends less time acquiring tandem mass spectra and the lack of alternating MS and MS/MS scans improves the ability to compare analyses. Later, ion chromatograms from an LC-MS system were used to identify differences between samples including complex mixtures such as digested serum with reasonable variation in the analyses (Wang et al., 2003). Wiener et al. (2004) used replicate LC-MS analyses to develop statistically significant differential displays of peptides. These approaches divide the comparison and identification processes into first identifying chromatographic and ion differences and then identifying the peptides responsible for the differences. To reduce comparison errors and ambiguities between samples, chromatographic peak alignment is increasingly used (Bylund et al., 2002; Maynard et al., 2004; Wiener et al., 2004; Wong et al., 2005; Katajamaa and Oresic, 2005; Zhang et al., 2005; Katajamaa et al., 2006).

By using the numbers of tandem mass spectra obtained for each protein or "spectral counting" as a surrogate for protein abundance in a mixture, Liu et al. (2004) demonstrated that "spectral counts" correlated linearly with protein abundance in a mixture within over two orders of magnitude. Because of the more complex nature of the LC/LC method and the alternating acquisition of mass spectra and tandem mass spectra, chromatographic alignment

is far more complicated than using LC-MS, and therefore, data are most often analyzed from the perspective of tandem mass spectra and identified proteins. Two issues with the use of LC/LC/MS/MS analyses to compare samples are the normalization of spectral counting data and the identification of differences between samples.

In the present study, we analyzed how well selected univariate and multivariate statistical/pattern recognition approaches can pinpoint protein markers, added at different concentrations to complex protein mixtures (yeast lysates), using spectral counting data. Different combinations of normalization/feature selection methods were applied and the combination that performed best on our dataset was identified by means of two approaches. The first ranked each protein by a statistical score according to which spiked markers were expected to rank highest. The second method relied on the SVM leave-one-out (LOO) cross-validation and the Vapnik-Chervonenkis (VC) confidence; briefly, these are quantifiers that allow the estimation of how well a classifier is to categorize unseen samples (Vapnik, 1995).

EXPERIMENTAL

MudPIT spectral count acquisition from yeast lysate with spiked proteins

Four aliquots of 400 µg of a soluble yeast total cell lysate were mixed with Bio-Rad SDS-PAGE low-range weight standards containing phosphorylase b, serum albumin, ovalbumin, lysozyme, carbonic anhydrase, and trypsin inhibitor at relative levels of 25, 2.5, 1.25, and 0.25% of the final mixtures' total weight, respectively. Each sample was sequentially digested, under the same conditions, with endoproteinase Lys-C and trypsin (Washburn et al., 2001). Approximately 70 µg of the digested peptide mixture was loaded onto a biphasic (strong cation exchange/reversed phase) capillary column and washed with a buffer containing 5% acetonitrile, 0.1% formic acid diluted in DDI water. The two-dimensional LC (LC/LC) separation and tandem MS (MS/MS) conditions were as described by Washburn et al. (2001). The flow rate at the tip of the biphasic column was 300 nL/min when the mobile phase composition was 95% H_2O, 5% acetonitrile, and 0.1% formic acid. The ion trap MS, Finnigan LCQ Deca (Thermo Electron, Woburn, MA, USA), was set to the data-dependent acquisition mode with dynamic exclusion turned on. One MS survey scan was followed by four MS/MS scans. Each aliquot of the digested yeast cell lysate was analyzed three times. The data sets were searched using a modified version of the *Pep_Prob* algorithm (Sadygov and Yates III, 2003) against a database combining yeast and human protein sequences, and the results were post-processed by DTASelect (Tabb et al., 2002). The sequences of the added markers and some common protein contaminants (e.g., keratin) were added to the database.

Generation of the three testing conditions

All computations in this study were performed using PatternLab for proteomics, available at http://pcarvalho.com/patternlab for academic use; its source code is also available upon request.

Firstly, PatternLab generated an index file listing all the proteins (features) identified in all the MudPIT assays. This index assigns a unique protein index number (PIN) to each feature. Secondly, all experimental data from the DTASelect files were combined into a single sparse matrix; this format is more suitable for feature selection. Each row of this matrix is relative to one MudPIT assay and gives the spectral count identified for each PIN in that assay.

Thus, for example, the row "1:3 2:5 3:6" specifies an assay having spectral count values of 3, 5, and 6 for PINs 1, 2, and 3, respectively; all other PINs are understood to have a value of 0. The sparse matrix generated for this study had 15 rows, obtained from 15 MudPIT runs with different percentages of protein markers added to the yeast lysate (4 runs with added markers representing 25% of the total protein content, 4 with 2.5%, 3 with 1.25%, and 4 with 0.25%). We note that each row had approximately 1200 PINs and a total of 2181 PINs were detected among all 15 rows, showing that many proteins were not identified in all runs.

Three testing datasets were then generated using the matrix, each one being identical to all others except for a class label introduced before each row. In the first test set (TSet1), the rows originating from the 25% protein spiking were labeled as +1 (positive) and all the others -1 (negative). In the second test set (TSet2), the 25% and the 2.5% matrix rows were labeled as +1 and the rest as -1. In the third (TSet3), the rows resulting from the 0.25% spiking were labeled as -1, the others as +1. The aim of such class labeling was to create 3 testing conditions for us to later compare the positively and negatively labeled rows in each testing dataset and verify whether the added proteins having different concentrations could be pinpointed. Figure 1 summarizes our methodology.

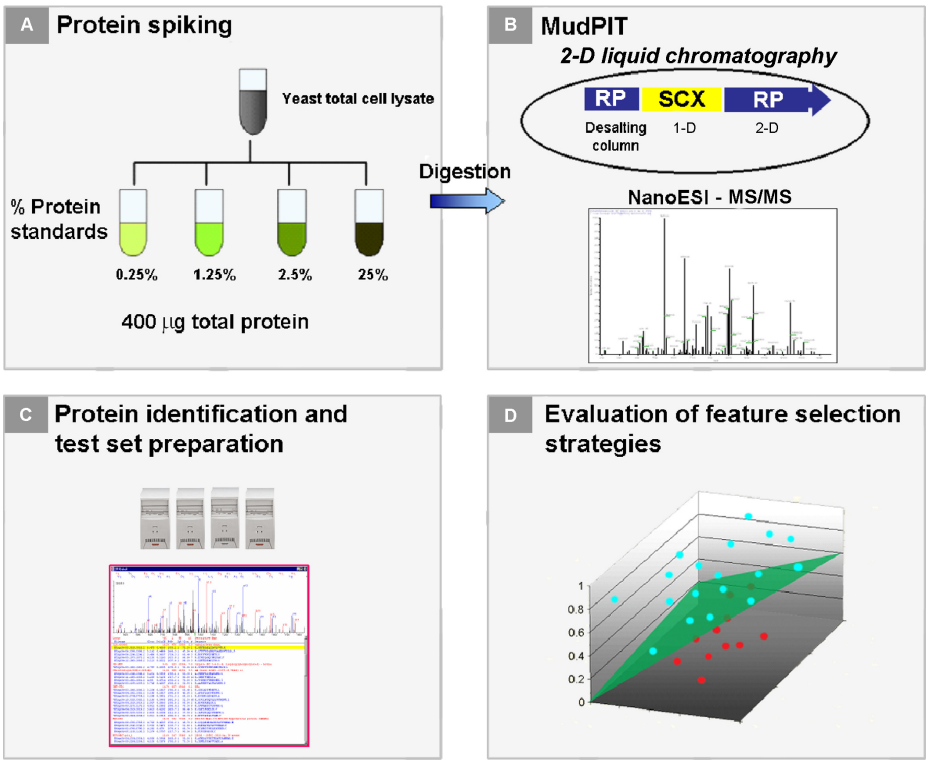

Figure 1. Protein markers were added at different concentrations to 15 yeast total cell lysate samples (A). Each lysate was analyzed by MudPIT (B) and protein identification carried out by *Pep_Prob* (C) and post-processed by DTASelect. Three different test sets were then generated. Combinations of normalization/feature selection methods were used to search for the added protein markers with different concentrations in each test set (D). RP = reverse phase material; SCX = strong cation exchange material; MS = mass spectrometry.

CALCULATION

Normalization methods evaluated in this study

For this study, we evaluated the following normalization strategies: total signal (TS), Z-normalization (Z), a hybrid normalization obtained by TS followed by Z (TS→Z), and log preprocessing.

Normalization by total spectral counting (total signal or TS)

Let SC_{ji} be the spectral count associated with PIN i in row j. The total spectral count (TSC) of row j is

$$TSC_j = \sum_i SC_{ji} \,. \qquad \text{(Equation 1)}$$

The normalization by TS of row j is obtained by performing

$$SC_{ji} \leftarrow \frac{SC_{ji}}{TSC_j} \qquad \text{(Equation 2)}$$

for all i.

Z-normalization

The Z-normalization has been widely adopted in microarray studies (Cheadle et al., 2003). For PIN i, let μ_i be the mean SC_{ji} over all j, and similarly σ_i the standard deviation. Normalization is achieved by performing

$$SC_{ji} \leftarrow \frac{SC_{ji} - \mu_i}{\sigma_i} \qquad \text{(Equation 3)}$$

for all j. The mean of the resulting SC_{ji} over all j is then zero and the standard deviation is 1. We note that Z is carried out over each matrix column while TS is performed on each matrix row.

Hybrid normalization (TS→Z)

This is obtained by TS followed by Z.

Log preprocessing

Taking the logarithm of the spectral count, data were also evaluated as a preprocessing step before the above normalization steps:

$$SC_{ji} \leftarrow \ln(SC_{ji}) \quad \text{(Equation 4)}$$

Our aim was to increase the signal of the PINs with low spectral counts with respect to the "highly abundant" PINs.

Feature selection/ranking methods evaluated in this study

For this study, we evaluated the Golub correlation coefficient, the Student *t*-test, a method we call forward-SVM (SVM-F), and SVM-RFE. All computations were carried out using PatternLab.

Golub index

For PIN *i*, the Golub index (GI; Golub et al., 1999) is defined by

$$GI_i = \frac{\mu_i^+ - \mu_i^-}{\sigma_i^+ + \sigma_i^-}, \quad \text{(Equation 5)}$$

where μ_i^+, μ_i^-, σ_i^+, and σ_i^- are the means and standard deviations of the data in column *i* restricted to the positive (+) or negative (-) class. The larger a positive GI_i the stronger the PIN's correlation will be with the positive class, whereas the smaller a negative GI_i the stronger the correlation with the negative class. For our goal of feature ranking, we simply took absolute values.

Student *t*-test

The score used for the Student *t*-test is given by

$$T_i = \frac{\mu_i^+ - \mu_i^-}{\sqrt{\frac{(n_i^+ - 1)s_i^+ + (n_i^- - 1)s_i^-}{n_i^+ + n_i^- - 2}\left(\frac{1}{n_i^+} + \frac{1}{n_i^-}\right)}}, \quad \text{(Equation 6)}$$

where each n_i is of the number of samples restricted to column *i* and to the positive (+) or negative (-) class, and each s_i is the corresponding variance. For our goal of feature ranking, we simply took absolute values.

Support vector machine

SVMs constitute a supervised learning method based on statistical learning theory and the principle of structural risk minimization (Vapnik, 1995). SVMs have been successfully used in a number of bioinformatics applications, including the prediction of

protein folds (Saha and Raghava, 2006), siRNA functionality (Teramoto et al., 2005), rRNA, DNA, and DNA-binding proteins (Yu et al., 2006), and the prediction of personalized genetic marker panels (Carvalho et al., 2006). An SVM model is evaluated using the most informative patterns in the data (the so-called support vectors) and is capable of separating two classes by finding an optimal hyperplane of maximum margin between the corresponding data.

Briefly, in the linearly separable case the SVM approach consists of finding a vector w in the feature space and a scalar b such that the hyperplane $\langle w, x\rangle + b$ can be used to decide the class, + or -, of input vector x (respectively if $\langle w, x\rangle + b \geq 0$ or $\langle w, x\rangle + b < 0$). During the training phase, the model's compromise between the empirical risk and its own complexity (related to its generalization capacity) is controlled by a penalty parameter C, a positive constant. We refer the reader to Vapnik's book for further details of the SVM approach, including how to obtain w and b from the training dataset (Vapnik, 1995). To carry out SVM modeling, PatternLab makes use of SVM*light* (Joachims, 1998).

SVM-F

SVM-F feature ranking is performed on the SVM model of the whole training set. If w is the corresponding vector in the feature space and w_i is the coordinate of w that corresponds to PIN i, then SVM-F ranks features in non-increasing order of w_i^2. Clearly, the lowest ranking PINs influence the hyperplane the least. SVM-F's output consists of the PINs ordered and listed side by side with their ranking scores.

SVM-RFE

SVM-RFE consists of recursively applying SVM-F to a succession of SVM models. The first of these corresponds to the whole training set; for $k > 1$, the k^{th} SVM model corresponds to the previously used training set after the removal of all entries that refer to the least-ranking PIN (according to SVM-F). The SVM models are then built on successively lower-dimensional spaces. Termination occurs when a desired dimensionality is reached or some other criterion is met. Since features are removed one at a time, an importance ranking can also be established.

Evaluation of combined normalization and feature-ranking methods

Combinations of the methods described were used to verify whether the added proteins could be pinpointed when comparing mixtures spiked with markers at different concentrations. In the ideal case, the four added proteins should achieve the top feature ranks. The ranks of the added proteins are listed in Tables 1 and 2 for the various method combinations and concentration comparisons. We used $C = 100$ for SVM training, following Guyon et al. (2002). The tables also show, in each case, a penalty score (Pscore) used to evaluate each method. This score plus one is the logarithm to the base 10 of the summed ranks of the four markers. Clearly, the ideal ranks yield a (minimum) Pscore of 0. Figure 2 plots the performance of each combination of normalization and feature ranking strategy.

Table 1. Normalization and feature selection results ($C = 100$).

No Log preprocessing																
	TS				Z				TS→Z				UD			
	GI	SVM-F	*t*-test	SVM-RFE	GI	SVM-F	*t*-test	SVM-RFE	GI	SVM-F	*t*-test	SVM-RFE	GI	SVM-F	*t*-test	SVM-RFE
TSet1																
PHS2	3	3	2	342	2	4	3	6	3	1	5	4	2	3	1	1242
ALB	1	1	18	340	1	2	1	7	1	2	132	1	1	1	2	469
CAH	4	4	187	343	17	3	25	2	4	4	301	5	18	4	9	1261
ITRA	2	2	30	341	3	1	2	1	2	3	128	2	3	2	7	476
Pscore	0	0	1.37	2.14	0.36	0	0.49	0.20	0	0	1.75	0.08	0.38	0.04	0.28	2.54
TSet2																
PHS2	2	5	1	7	4	4	10	14	2	6	1	56	4	4	7	20
ALB	1	1	2	1	1	2	1	1	1	3	2	18	1	1	2	14
CAH	4	3	4	4	2	1	2	4	4	1	5	23	2	3	1	16
ITRA	3	2	3	2	3	3	3	12	3	2	4	22	3	2	3	15
Pscore	0	0.04	0	0.15	0	0	0.2	0.49	0	0.08	0.08	1.08	0	0	0.11	0.81
TSet3																
PHS2	352	4	5	173	10	4	111	18	352	7	7	15	9	4	68	315
ALB	348	1	2	8	2	1	2	2	348	5	3	54	2	1	8	313
CAH	357	3	1	6	1	2	1	1	357	3	5	61	1	3	2	312
ITRA	361	2	4	12	3	3	5	17	361	6	4	43	3	2	10	314
Pscore	2.15	0	0.08	1.30	0.20	0	1.1	0.58	2.15	0.32	0.28	1.24	0.18	0	0.94	2.10
PSum	2.15	0.04	0.93	3.59	0.56	0	0.74	1.27	2.15	2.4	1.30	2.4	0.56	0.04	0.60	5.45

The first column lists the added proteins we tracked: phosphorylase b (PHS2), serum albumin (ALB), carbonic anhydrase (CAH), and trypsin inhibitor (ITRA). We give results on three normalization methods, in addition to results on the "unnormalized" data (UD): total signal (TS), Z-normalization (Z) and TS followed by Z. GI, *t*-test, SVM-F and SVM-RFE stand for Golub's index, Student *t*-test , forward support vector machine, and SVM-recursive feature elimination, the four feature selection methods. Data are presented for all different test set analyses (refer to the end of the section "Generation of the three testing conditions"). Each number indicates the rank of the protein markers among the various other proteins present in the yeast lysate. To compare the method combinations, we used a penalty score (Pscore) that is calculated as log (sum of the added protein ranks) -1. Here we used 10 as the log's base; thus, for a perfect score the ranks add up to 10 (4 + 3 + 2 + 1) yielding the Pscore as 0. PSum is the sum of the Pscores for a given method and is used to quickly search for which performed best. TSet 1, TSet 2 and TSet 3 are first, second and third test sets, respectively.

Table 2. Normalization and feature selection results (C = 100) after log preprocessing.

	Log preprocessing															
	TS				Z				TSC→Z				UD			
	GI	SVM-F	*t*-test	SVM-RFE	GI	SVM-F	*t*-test	SVM-RFE	GI	SVM-F	*t*-test	SVM-RFE	GI	SVM-F	*t*-test	SVM-RFE
								TSet1								
PHS2	59	359	44	522	31	1	5	7	59	228	305	194	31	1	2	6
ALB	7	75	159	373	58	2	132	6	7	2	389	11	58	3	17	8
CAH	6	75	111	372	252	4	301	22	6	1	426	5	252	2	186	11
ITRA	8	101	122	375	87	3	128	14	8	3	398	8	87	4	29	10
Pscore	0.90	1.79	1.64	2.22	1.63	0	1.75	0.69	0.90	1.37	2.18	1.34	1.63	0	1.40	0.54
								TSet2								
PHS2	1017	351	7	432	2	1	1	1	1017	234	1	312	2	1	1	1
ALB	9	80	29	25	1	4	2	2	9	2	5	7	1	3	2	3
CAH	6	75	16	38	5	3	5	15	6	1	295	1	5	2	4	8
ITRA	30	88	22	32	3	2	3	7	30	3	3	6	3	4	3	7
Pscore	2.03	1.77	0.87	1.72	0.04	0	0.04	0.40	2.03	1.38	1.48	1.51	0.04	0	0	0.28
								TSet3								
PHS2	2088	407	1	512	4	2	6	3	2088	247	2	354	4	1	5	9
ALB	892	91	12	83	2	3	3	7	892	2	4	10	2	3	2	8
CAH	664	82	2	56	1	4	5	1	664	1	1	2	1	2	1	1
ITRA	1352	100	8	113	3	1	4	2	1352	3	3	11	3	4	4	7
Pscore	2.70	1.83	0.36	1.88	0	0	0.26	0.11	2.72	1.38	0	1.58	0	0	0.08	0.40
PSum	5.63	5.39	1.25	5.82	1.67	0	1.30	1.2	5.65	4.13	1.79	4.43	1.67	0	0.93	1.22

For abbreviations, see legend to Table 1. Log_e was used as a preprocessing step before qualifying the feature ranking methods.

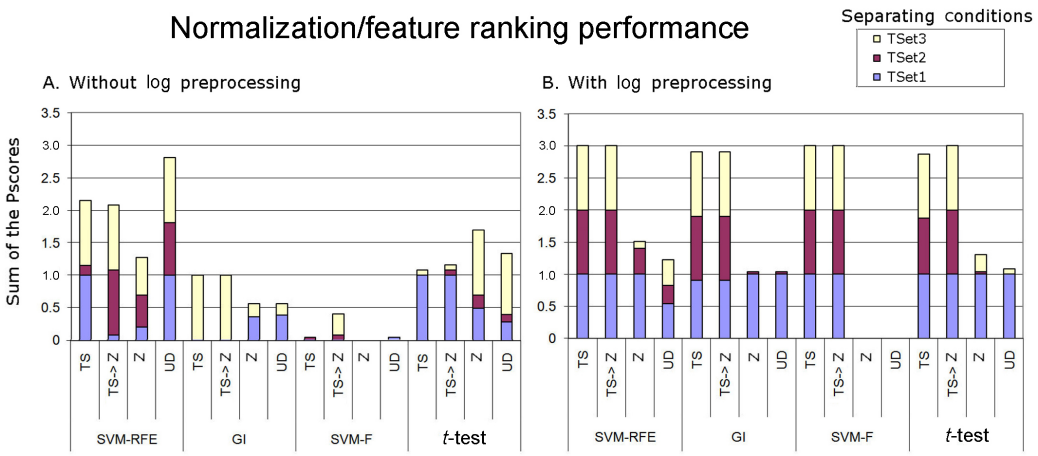

Figure 2. Sum of penalty scores (Pscores) calculated for each combination of normalization/feature selection method when comparing the different spike concentrations (legend), with (B) and without (A) log preprocessing. Lower bars indicate better performance. The bar heights were limited to 4. We recall that the Pscore is calculated by obtaining the Log_{10} of the sum of the ranks and subtracting 1. Note that SVM-F with and without log preprocessing obtains at least one perfect score. UD stands for "unnormalized" data. For abbreviations, see legend to Table 1.

Evaluation of the normalization methods

By using only the spectral counts of the added proteins, SVM models were also calculated varying the *C* parameter from 2 to 100 with a step of 2 for all normalization methods. The *C*s that achieved a minimum LOO error or VC confidence were recorded. In either case, the LOO error, the VC confidence, and the number of support vectors of the model were also recorded (Table 3). We note that LOO error and VC confidence are respectively ways of measuring a model's empirical risk (the error within the dataset) and how much may be added to that risk as the model is applied to a new dataset (generalization capacity).

The LOO technique consists of removing one example from the training set, computing the decision function with the remaining training data and then testing it on the removed example. In this fashion one tests all examples of the training data and measures the fraction of errors over the total number of training examples.

The model's VC confidence has roots in statistical learning theory (Vapnik, 1995) and is given by

$$\text{VC confidence} = \sqrt{\frac{h(\ln(2l/h)+1) - \ln(\eta/4)}{l}}, \qquad \text{(Equation 7)}$$

were h is the VC dimension of the model's feature space, l is the number of training samples and $1\text{-}\eta$ is the probability that the VC confidence is indeed the maximum additional

error to the empirical risk as new datasets are presented to the model. We used $\eta = 0.05$ throughout. We recall that, given an SVM model, the VC dimension is a function of the separating margin between classes and the smallest radius of the hypersphere that encompasses all input vectors.

Table 3. Linear support vector machine (SVM) separability analysis.

	No log preprocessing				Log preprocessing			
Norm.	TS	Z	TS→Z	UD	TS	Z	TS→Z	UD
				TSet1				
C for Min VC	2	2	2	2	2	2	2	2
C for min LOO	86	2	2	2	2	2	2	2
VC-LOO	0.27	0	0	0	0	0	0	0
mLOO	0	0	0	0	0	0	0	0
VC-Conf-mVC	0.624	1.333	1.027	1.871	2.301	1.949	1.501	2.503
VC-LOO-SV	8	3	2	2	3	4	2	3
mLOO-SV	8	3	2	2	3	4	2	3
				TSet2				
C for Min VC	2	2	2	2	2	2	2	2
C for min LOO	2	2	2	2	54	2	2	2
VC-LOO	0.47	0	0.27	0	0.467	0	0.20	0
mLOO	0.47	0	0.27	0	0.333	0	0.20	0
VC-Conf-mVC	0.624	1.278	0.775	2.013	>2.753	1.239	1.641	2.431
VC-LOO-SV	8	4	8	3	8	4	9	2
mLOO-SV	8	4	8	3	8	4	9	2
				TSet3				
C for Min VC	4	2	4	2	6	2	2	2
C for min LOO	4	2	4	2	6	2	4	2
VC-LOO	0.27	0	0.27	0	0.200	0	0.267	0
mLOO	0.27	0	0.27	0	0.200	0	0.200	0
VC-Conf-mVC	0.624	1.841	0.633	1.265	1.470	1.280	1.625	1.673
VC-LOO-SV	9	4	10	2	8	2	8	2
mLOO-SV	9	4	10	2	8	2	8	2

C for Min VC and **C for min LOO** represent the *C* values used during the SVM training that achieved the minimum Vapnik-Chervonenkis (VC) confidence and the minimum leave-one-out (LOO) error, respectively. **VC-LOO** and the **mLOO** are the LOO errors obtained when **C for Min VC** and **C for min LOO** are used during the SVM training phase. **VC-Conf-mVC** represents the model's VC confidence when the model was trained with **C for min LOO**. **VC-LOO-SV** and the **mLOO-SV** represent the number of support vectors contained in the classification model when trained with **C for Min VC** and **C for min LOO**, respectively. For other abbreviations, see legend to Table 1.

Predicting how many proteins were added

Feature ranking can be combined with methods that predict how many features are significant. Here, predicting the number of features is equivalent to estimating how many proteins were added. All feature ranking methods we used output a two-column list having features (PINs) ordered by their ranks in the first column and the method's score for each PIN in the second column. The number of added proteins was estimated by locating, in this output list, the two consecutive rows that presented the greatest difference in score values. The number of features was then computed by counting how many features have scores above or equal to this gap's upper limit.

RESULTS AND DISCUSSION

Evaluation of the feature selection/ranking methods

An efficient feature ranking criterion should select the features that best contribute to a learning machine's ability to "separate" data, reduce pattern recognition costs, and make the model less prone to overfitting. Translational studies usually possess a limited number of samples and have high dimensionality (many features), making feature selection and evaluation of the generalization capacity imperative steps. By spiking yeast lysates with proteins and then detecting them, we perform a proof of principle of the potential of using spectral counts and SVMs to identify differences and perform classification in proteomic profiles.

In our hands, for the yeast MudPIT spectral count dataset, both Z-normalization, with and without log preprocessing, and the use of "unnormalized" data with log preprocessing followed by SVM-F achieved a perfect score, pinpointing all added proteins for all configurations over the 10^2 dynamic range tested. These results are shown in Tables 1 and 2, and Figure 2.

Overall, the greatest difficulties were in locating the added markers in TSet1. We hypothesize that this originates from limitations in both the feature selection methods and the experimental procedure used. From the machine learning perspective, according to Cover and Van Campenhout (1977), no non-exhaustive sequential feature selection procedure is guaranteed to find the optimal feature subset or list the ordering of the error probabilities. We do not use exhaustive feature searching, since the number of subset possibilities grows exponentially with the number of features; this method quickly becomes unfeasible, even for a moderate number of features. Less abundant proteins are not identified for every MudPIT analysis, generating a bias toward the acquisition of the more abundant peptide ions. Thus, less abundant proteins are identified by fewer peptides, and their identifications can sometimes be suppressed by peptides from more abundant proteins. Liu et al. (2004) addressed the randomness of protein identification by MudPIT for complex mixtures. The rows originating from TSet1 show that fewer PINs were identified during these runs (~800), contrasting with the ~1200 PINs from the other runs. This lack of PINs may have driven the SVM-RFE toward an "undesired direction" while recursively eliminating the features. During the RFE computation and before narrowing down to ~600 features, the weights of the normal vector (w) still included the added proteins among the most important features.

Although we successfully identified the added proteins, we believe our methods could develop into variants that could perform better for datasets of a different nature. The methods we employed are deterministic, in the sense that they quickly narrow down to what may be only locally optimal solutions. The quest for the global optimum in high-dimensional feature spaces still remains a challenge for pattern recognition. Distributed computing, coupled with algorithms that can efficiently rake the feature space (genetic algorithms (Shah and Kusiak, 2004; Link et al., 1999), swarms (Guo et al., 2004), etc.), holds promises for proteomics of mining datasets more complex than the ones we addressed.

Evaluation of the normalization methods regarding dataset "separability"

Given that more than one method is able to select the added proteins, which one is best? Since added markers exist in different concentrations in each class and since spectral

counts correlate with protein abundance, there should be a linear function capable of separating the input vectors containing only the spectral count information of the added proteins. To further evaluate the generalization capacity of the model, we used the VC confidence.

Both Z and log preprocessed data allowed SVM-F to correctly select the added proteins and yielded a 0% LOO error for all spiking configurations (Table 3). VC confidence shows that TSet1 and TSet2 normalized by Z led to a greater capacity than TSet3, thus here the lower concentrations made it harder for Z preprocessing. On the other hand, the log preprocessed data separated better for the lower concentrations, probably because of the nature of the log function which discriminates lower values better than larger values.

In our results, the feature selection methods applied to "unnormalized" data achieved good Pscores. We hypothesize that this happened because the datasets were similar in the sense that the background proteins were technical replicates (thus easily reproducible). Had the yeast proteins been more variable, then it is possible that the normalization methods would become critically important. Further research is needed on this.

Predicting the number of added proteins

Overall, according to our benchmarking strategy, Z-normalization followed by SVM-F was the method that obtained a perfect score for the yeast MudPIT dataset. The method previously described to predict the number of added markers was applied to the Z/SVM-F results, and it correctly identified the number of added markers as being 4 for all three possibilities of spiked-lysate separation (TSet1 through 3).

CONCLUSIONS

In this study, we set out to address the question of whether the data from spectral counts can be normalized and then classified using pattern recognition techniques. The above results indicate that Z followed by SVM-F applied to the yeast MudPIT spectral count dataset is an effective method for finding differences in this type of data. The methodology described was also capable of correctly identifying how many markers were added to the lysate. It is expected that the presented method should perform satisfactorily for other experiments where data are similarly acquired.

The identification of trustworthy marker proteins is not an easy task, since mass spectrometry-based proteomics is still in development and spectral counting effectiveness can vary with the experimental setup, including mass spectrometry type and data-dependent analysis configuration. Here, combinations of normalization and feature selection strategies were validated on a controlled (spiked) but realistic (yeast lysate) experiment, which is therefore verifiable. Our results indicate that even in "simple" scenarios where the spike concentrations can be considered relatively high, the data can still play tricks on well-founded feature selection methods. This is due to the dataset's high dimensionality, sparseness, and lack of a known *a priori* probability distribution. For even more complex scenarios, the searched markers could be present in extremely low concentrations when compared to the absolute concentrations. One of the existing strategies to reduce complexity is to isolate sub-proteomes; however, these separations are many times not straightforward to be carried out while disturbing protein content only minimally and remain a challenge.

We have also demonstrated the importance of evaluating computational strategies for proteomics studies to verify which one best suits the experiment at hand before drawing conclusions when dealing with complex datasets. As shown by our results, the application of SVM-RFE in our spectral count yeast dataset could lead to false conclusions. This shows that pattern recognition methods can perform differently with datasets of distinct natures, strengthening the idea that there is no "one suits all" method.

ACKNOWLEDGMENTS

Research supported by funds from the National Institutes of Health (P41 RR11823-10, 5R01 MH067880, and U19 AI063603-02), CNPq, CAPES, a FAPERJ BBP grant, and the Genesis Molecular Biology Laboratory. The authors thank Dr. Hongbin Liu for sharing MudPIT data (Liu et al., 2004).

REFERENCES

Badr G and Oommen BJ (2006). On optimizing syntactic pattern recognition using tries and AI-based heuristic-search strategies. *IEEE Trans. Syst. Man. Cybern. B. Cybern.* 36: 611-622.

Bylund D, Danielsson R, Malmquist G and Markides KE (2002). Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography-mass spectrometry data. *J. Chromatogr. A* 961: 237-244.

Carlson JM, Chakravarty A and Gross RH (2006). BEAM: a beam search algorithm for the identification of cis-regulatory elements in groups of genes. *J. Comput. Biol.* 13: 686-701.

Carvalho PC, Freitas SS, Lima AB, Barros M, et al. (2006). Personalized diagnosis by cached solutions with hypertension as a study model. *Genet. Mol. Res*. 5: 856-867.

Carvalho PC, Carvalho MG, Degrave W, Lilla S, et al. (2007). Differential protein expression patterns obtained by mass spectrometry can aid in the diagnosis of Hodgkin's disease. *J. Exp. Ther. Oncol.* 6: 137-145.

Cheadle C, Vawter MP, Freed WJ and Becker KG (2003). Analysis of microarray data using Z score transformation. *J. Mol. Diagn.* 5: 73-81.

Cover TM and Van Campenhout JM (1977). On the possible orderings in the measurement selection problem. *IEEE Trans. Syst. Man. Cybern.* 7: 651-661.

Golub TR, Slonim DK, Tamayo P, Huard C, et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286: 531-537.

Guo CX, Hu JS, Ye B and Cao YJ (2004). Swarm intelligence for mixed-variable design optimization. *J. Zhejiang. Univ. Sci.* 5: 851-860.

Guyon I, Weston J, Barnhill S and Vapnik V (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn.* 46: 389-442.

Jessani N, Niessen S, Wei BQ, Nicolau M, et al. (2005). A streamlined platform for high-content functional proteomics of primary human specimens. *Nat. Methods* 2: 691-697.

Joachims T (1998). Making large-scale support vector machine learning practical. In: Advances in Kernel methods: support vector machines (Scholkopf B, Burges C and Smola A, eds.). MIT Press, Cambridge, 169-185.

Katajamaa M and Oresic M (2005). Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics* 6: 179.

Katajamaa M, Miettinen J and Oresic M (2006). MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* 22: 634-636.

Kolakowska A and Malina W (2005). Fisher sequential classifiers. *IEEE Trans. Syst. Man. Cybern. B. Cybern.* 35: 988-998.

Link AJ, Eng J, Schieltz DM, Carmack E, et al. (1999). Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* 17: 676-682.

Liu H, Sadygov RG and Yates JR III (2004). A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* 76: 4193-4201.

Mattie MD, Benz CC, Bowers J, Sensinger K, et al. (2006). Optimized high-throughput microRNA expression profiling

provides novel biomarker assessment of clinical prostate and breast cancer biopsies. *Mol. Cancer* 5: 24.
Maynard DM, Masuda J, Yang X, Kowalak JA, et al. (2004). Characterizing complex peptide mixtures using a multi-dimensional liquid chromatography-mass spectrometry system: *Saccharomyces cerevisiae* as a model system. *J. Chromatogr. B. Analyt. Technol. Biomed. Life Sci.* 810: 69-76.
Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, et al. (2002). Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 359: 572-577.
Polisetty PK, Voit EO and Gatzke EP (2006). Identification of metabolic system parameters using global optimization methods. *Theor. Biol. Med. Model.* 3: 4.
Sadygov RG and Yates JR III (2003). A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal. Chem.* 75: 3792-3798.
Saha S and Raghava GP (2006). VICMpred: an SVM-based method for the prediction of functional proteins of Gram-negative bacteria using amino acid patterns and composition. *Genomics Proteomics Bioinformatics* 4: 42-47.
Shah SC and Kusiak A (2004). Data mining and genetic algorithm based gene/SNP selection. *Artif. Intell. Med.* 31: 183-196.
Tabb DL, McDonald WH and Yates JR III (2002). DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* 1: 21-26.
Teramoto R, Aoki M, Kimura T and Kanaoka M (2005). Prediction of siRNA functionality using generalized string kernel and support vector machine. *FEBS Lett.* 579: 2878-2882.
Unlu M, Morgan ME and Minden JS (1997). Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis* 18: 2071-2077.
Vapnik VN (1995). The nature of statistical learning theory. Springer-Verlag New York Inc., New York.
Wang W, Zhou H, Lin H, Roy S, et al. (2003). Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal. Chem.* 75: 4818-4826.
Washburn MP, Wolters D and Yates JR III (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* 19: 242-247.
Wiener MC, Sachs JR, Deyanova EG and Yates NA (2004). Differential mass spectrometry: a label-free LC-MS method for finding significant differences in complex peptide and protein mixtures. *Anal. Chem.* 76: 6085-6096.
Wong JW, Cagney G and Cartwright HM (2005). SpecAlign - processing and alignment of mass spectra datasets. *Bioinformatics* 21: 2088-2090.
Yu X, Cao J, Cai Y, Shi T, et al. (2006). Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines. *J. Theor. Biol.* 240: 175-184.
Zhang X, Asara JM, Adamec J, Ouzzani M, et al. (2005). Data pre-processing in liquid chromatography-mass spectrometry-based proteomics. *Bioinformatics* 21: 4054-4059.