JOURNAL OF PROTEOMICS 75 (2012) 2269–2274

Available online at www.sciencedirect.com

SciVerse ScienceDirect

www.elsevier.com/locate/jprot

Technical Note

# Isoelectric point optimization using peptide descriptors and support vector machines

Yasset Perez-Riverol[a,d], Enrique Audain[b], Aleli Millan[a], Yassel Ramos[a], Aniel Sanchez[a], Juan Antonio Vizcaíno[d], Rui Wang[d], Markus Müller[c], Yoan J. Machado[b], Lazaro H. Betancourt[a], Luis J. González[a], Gabriel Padrón[a], Vladimir Besada[a,*]

[a]Department of Proteomics, Center for Genetic Engineering and Biotechnology, Ave 31 e/ 158 y 190, Cubanacán, Playa, Ciudad de la Habana, Cuba
[b]Department of Proteomics, Center of Molecular Immunology, Calle 15 esq. 216, Siboney, Playa, Ciudad de la Habana, Cuba
[c]Proteome Informatics Group, Swiss Institute of Bioinformatics, CMU - 1, rue Michel Servet CH-1211 Geneva, Switzerland
[d]EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

ARTICLE INFO

*Article history:*
Received 22 November 2011
Accepted 25 January 2012
Available online 3 February 2012

*Keywords:*
Isoelectric point
Support vector machine
Peptide descriptors

ABSTRACT

IPG (Immobilized pH Gradient) based separations are frequently used as the first step in shotgun proteomics methods; it yields an increase in both the dynamic range and resolution of peptide separation prior to the LC-MS analysis. Experimental isoelectric point (*pI*) values can improve peptide identifications in conjunction with MS/MS information. Thus, accurate estimation of the *pI* value based on the amino acid sequence becomes critical to perform these kinds of experiments. Nowadays, *pI* is commonly predicted using the charge-state model [1], and/or the cofactor algorithm [2]. However, none of these methods is capable of calculating the *pI* value for basic peptides accurately. In this manuscript, we present an new approach that can significant improve the *pI* estimation, by using Support Vector Machines (SVM)[3], an experimental amino acid descriptor taken from the AAIndex database [4] and the isoelectric point predicted by the charge-state model. Our results have shown a strong correlation ($R^2$=0.98) between the predicted and observed values, with a standard deviation of 0.32 pH units across the complete pH range.



Isoelectric point can be defined as the point in a titration curve at which the net surface charge of a protein or peptide equals to zero [5]. The technique of using isoelectric focusing (IEF), where molecules are separated on the basis of their isoelectric points, for the separation of protein mixtures has been widely employed. Electrophoresis-based separation of peptides in both free-flow and gel systems (along with the subsequent *pI* calculations) has been adapted to a wide variety of proteomics platforms as the separation step, which reduces the complexity of the studied proteome [2,6,7]. In addition to the inherent high resolution (gel IPG-based approach) and dynamic range, combining the electrophoretic separation of peptides with MS/MS

*Abbreviations:* FDR, False Discovery Rate; IPG, Immobilized pH Gradient; RMSD, Root-mean-square deviation; SVM, Support Vector Machine.

* *Corresponding author at:* Center for Genetic Engineering and Biotechnology, Apartado 6162, POB 10600, La Habana, Cuba, Fax: +53 1573 271 6022.
E-mail address: vladimir.besada@cigb.edu.cu (V. Besada).

1874-3919/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.jprot.2012.01.029

provides an orthogonal (*pI*) analysis method for either database filtering or validation of the peptide identifications [8] in different workflows.

Current algorithms for estimating isoelectric points of peptides and proteins depends primarily on the model proposed by Bjellqvist and co-workers [1]. This model is based on the determination of the *pK* differences between closely related immobilines, by focusing the same sample in overlapping pH gradients. Some improvements in the methodology (especially in the determination of the *pK* values) have been published since [6,9]. As a alternative, Cargile and co-workers [2] has followed another approach: their algorithm accounts for the effect of adjacent amino acids ±3 residues away from a charged aspartic or glutamic acid, the effects on free C terminus, as well as applies a correction term to the corresponding *pK* values [2]. They also applied genetic optimization method to a 5000-peptide training set to derive the results, which have shown isoelectric point not only depends on individual amino acid, but also on the interactions between different amino acids present in the peptide sequence. The accuracy of the new *pI* values obtained with this method is close to the error associated with the manufacturer of the IPG strips (±0.03 *pI* units). However, the algorithm and the adjusted *pK* values were optimized only for the acid pH range (from 3.5 to 4.5), where most of peptides are well resolved.

In this manuscript, we introduce a new approach that improves the existing methods in the basic pH range (7–14 pH units), and can also be used in the acid range (from 3.5 to 7). It uses Support Vector Machines (SVM) as predictors, and takes into account both an experimental amino acid descriptor from the AAIndex database [4] and the isoelectric points predicted by the *Bjellqvist* model [1].

In recent years, there have been vast interests in studying Support Vector Machines (SVMs) approaches in the field of machine learning, this is due to their many appealing features and promising empirical performances. To date, SVMs have been applied successfully to a broad range of regression problems in proteomics, such as identification of protein cleavage sites, amino acid retention time and isoelectric point prediction [10–13]. In the case of applying SVMs to the prediction of

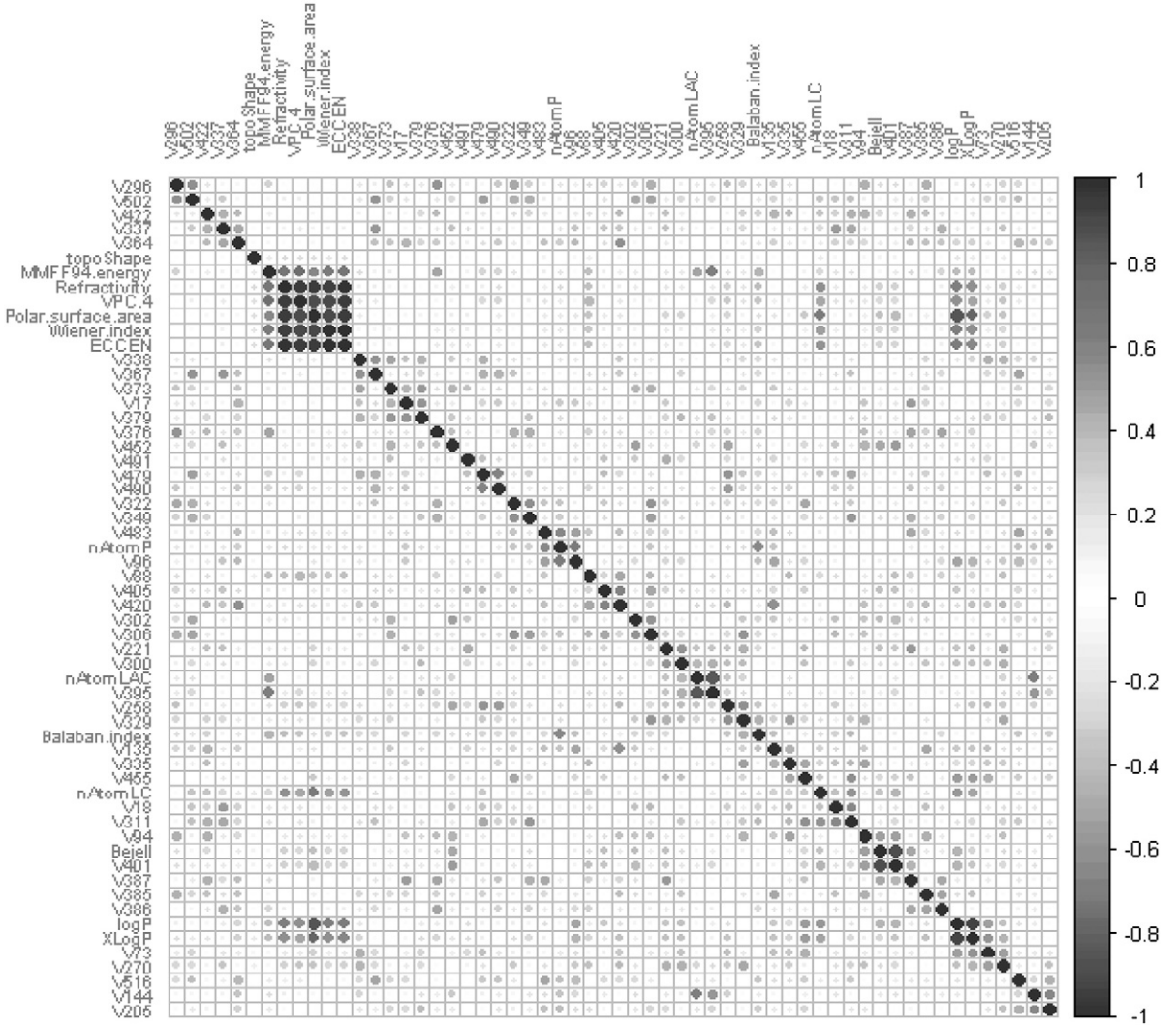


**Fig. 1 – Matrix correlation of the final features selected to train and tests the model. The top and left axes represent each property. The right gradient represents the correlations from –1 to 1. The matrix cells represent the correlation between a pair of descriptors.**

isoelectric point, a concise and meaningful encodings of the peptide properties are essential. These properties are mainly determined by the overall amino acid composition. Then, from the set of properties different feature selection algorithms must be applied to select the most prominent predictors. Finally, SVMs use a kernel function to encode distances between individual data points (peptides). The central idea is to map data *x* into a higher dimensional feature space *F* via a nonlinear mapping and then do linear regression in this space [14].

We have adopted the technique used by the ChemAxon (http://www.chemaxon.com) package. Several molecular descriptors (such as refractivity index, polarizability, surface area, LogP) were carefully studied. In addition, we included the physicochemical and biological properties of the amino acids from the AAIndex database [4]. AAindex is a database of numerical indices representing various physicochemical and biochemical properties of amino acids. For each peptide, all 543 existing descriptors were computed in AAIndex with the following mathematical expression: $P_D=(\sum A_D)/N_A$ , where the peptide AAindex descriptor ($P_D$) is the average of all amino acid AAindex descriptors ($A_D$). We also included the *pI* values predicted with the *Bjellqvist* [1] and *Cargile* [2] approaches as a peptide descriptor.

The new SVM based model to predict isoelectric point was trained, developed and tested on the peptides from a study carried out with *D. melanogaster* Kc167 cells, where an OFFGEL electrophoresis was performed as the first separation step. The peptide samples were then analysed on a LTQ-FT-ICR instrument equipped with a nanoelectrospray ion source (for more details about the study, see Supplementary Information) [8]. For the identification step, X!Tandem [15] and PeptideProphet [16] were used. Identified peptides were divided in two groups: one group consists of only highly reliable peptide identifications, identifications with PeptideProphet probability score higher than 0.97 (correspond to 0.01 FDR) and another group contains peptide identifications with PeptideProphet probability lower than 0.97. Peptides containing post-translational modifications (PTMs) were not considered. Furthermore, the redundant peptide identifications for each well were not eliminated in order to prevent overestimation in the model.

The first group with 7391 more reliable identifications was used to train, generate and test the model. After the calculation of all the previously described peptide descriptors, a data reduction step was performed to select the relevant descriptors that can describe dependably the property of interest (in this case, the *pI*).

In the first stage, we computed a correlation matrix on the predictors and then remove the subset of the problematic predictors (more correlated). Consequently, all the descriptors with pairwise correlations higher than 0.7 were removed (Fig. 1). However, one cluster, represented by five variables (refractivity, polar surface area, wiener index, topological shape, MMFF94 energy) was left in intentionally, on the grounds that previous reports have shown the correlation between isoelectric point and these molecular properties [10]. Finally, the feature select algorithm reduces the feature space from 555 to 44 descriptors.

The second stage is a feature selection step in combination with a SVM algorithm, which were written in R using the *caret* package [17]. The 7391 peptides were randomly partitioned into a training (75%) and test (25%) dataset to construct the SVM model. The feature selection part of the algorithm is based on simple-recursive backwards method. Let *S* be a sequence of ordered numbers which are candidate values for the number of predictors to retain ($S_1>S_2,\ldots$). After each iteration of the feature selection, the $S_i$ top ranked predictors is retained. In the end, the top $S_i$ predictors with the best performing $S_i$ values are used in the final model. The SVM model with a specific kernel function has been applied to evaluate the selected predictors and to generate the final model.

We have evaluated four different SVMs function kernels with automated sigma estimation using the *kernlab* R-package [18]: polynomial, lineal, exponential and radial (Table 1). The best results were obtained using the radial function, which has shown the Pearson correlation between the experimental and the theoretical *pI* values higher than ($R^2=0.98$). Furthermore, the RMSD (root-mean-square deviation) was 0.32 units, for the complete range of pH. The final model selects only two predictors to estimate accurately the isoelectric points of peptides, they are the isoelectric point predicted with the *Bjellqvist* algorithm and the experimental AAindex descriptor from *Zimmerman* [19]. The *Zimmerman* index is related to with the isoelectric point of individual amino acids. In contrast with previous results [10], the isoelectric points of peptides are not related to the polarity, the reactivity or the bulkiness of the molecule.

Our approach does not require huge computing power. All the calculations were performed in a standard-spec computer (Intel Core 2 Duo, 2 GB Ram). The time used to compute all peptide descriptors was 8.56 seconds (s), and the feature selection algorithm took 12 s. Training, predicting and generating the final model took 22 min in total.

A direct comparison of the previous developed algorithms and our SVM based approach is shown in Fig. 2. The overall correlation between the experimental and theoretical *pI* values was $R^2=0.91$ (adjacent algorithm), $R^2=0.96$ (charge-state), and $R^2=0.98$ (SVM algorithm). The standard deviation of the SVM method decreased to 0.32 pH units compared to 0.37 and 0.38 for charge-state and adjacent algorithms, respectively.

In general, small standard deviation in all fractions is observed. Particularly, the theoretical and experimental values are more correlated in the 3.0–4.0 *pH* range. This is due the number of peptide identifications presents in those fractions and the fact that the SVM algorithm is an optimization of the charge-state algorithm (*Bjellqvist*), by adding an extra experimental AAindex descriptor. Similarly, *the adjacent* algorithm

**Table 1 – The 4 kernel functions evaluated with the training set. The variable used to select the best kernel was the RMSD (root-mean-square deviation). (a) The number of variable predictors to generate the model. $R^2$ is the correlation between the experimental values and the theoretical.**

| Kernel function | Number of predictors[a] | RMSD | $R^2$ |
|---|---|---|---|
| Polynomial | 25 | 0.3387 | 0.97 |
| Lineal | 20 | 0.3866 | 0.96 |
| Exponential | 2 | 0.4 | 0.96 |
| Radial | 2 | 0.32 | 0.98 |

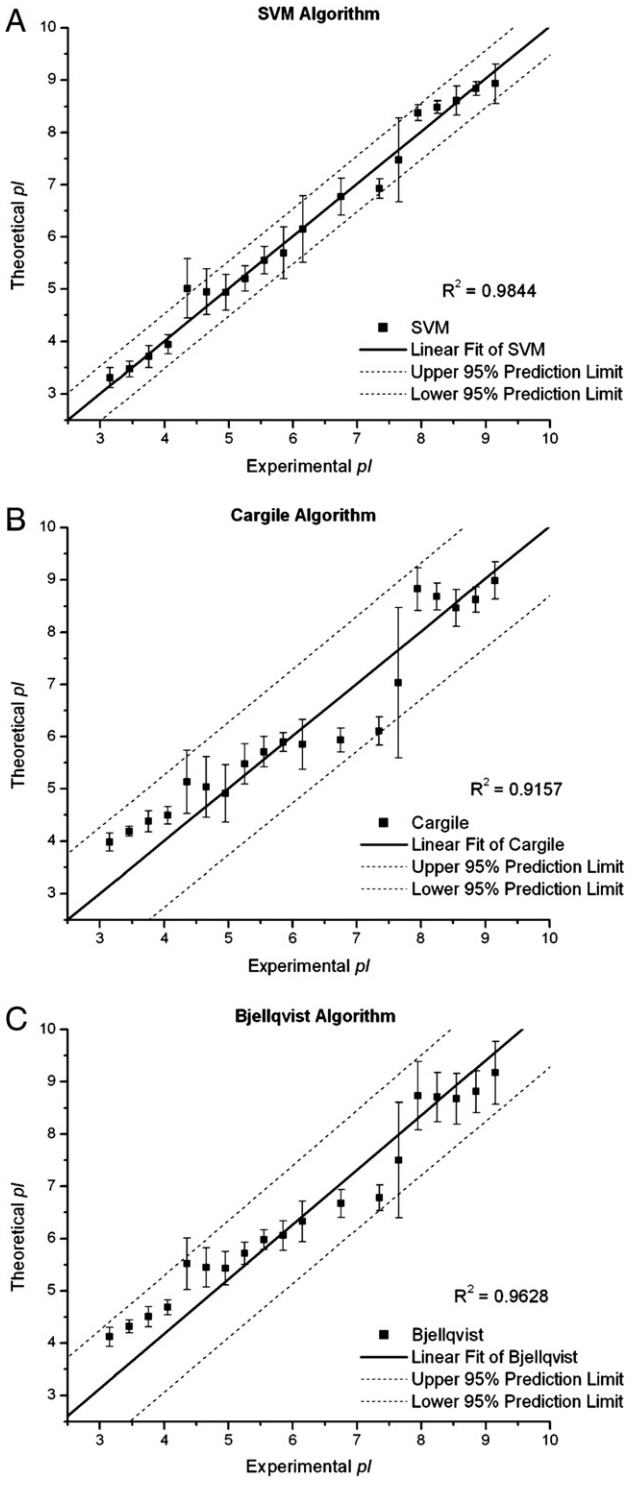


**Fig. 2 – Plot of experimental *vs* theoretical *pI* for our support vector machine algorithm (A), Cargile algorithm [2] (B) and charge-state algorithm from Bjellqvist [1] (C). The x axes correspond to the experimental isoelectric point range of 3–9 (24 fractions).**

(*Cargile*) showed a very good performance in the first four fractions from the acidic region (on the 3.5–4.5 *pH* range). This is also the envisaged result as the algorithm was originally trained on 5000 unique tryptic peptides separated on an 18-cm *pI* 3.5–4.5 IPG strip. The average of the standard deviation for the first five fractions for the SVM model, *the charge-state* and *the adjacent* algorithm was 0.26, 0.23 and 0.25 respectively.

The results were even better for the last five fractions, the most basic ones (7.65, 7.95, 8.25, 8.55, 9). In these fractions the average of the standard deviation (*stdv*) was 0.20, 0.52, 0.32 for the SVM model, the charge-state and the adjacent algorithm respectively (Fig. 2). The confidence interval (95% of confidence level) is better for the SVM prediction compares to the values from the *Cargile* and *Bjellqvist* methods. The use of SVM algorithms and machine learning techniques in general, give the possibility to find a new model to predict the isoelectric point given some background knowledge (reliable identifications) from all fractions [20–23]. Another published dataset was also used to demonstrate that the model can predict accurately across diverse dataset and experimental settings. We used the PeptideProphet dataset extracted from the Heller and cols [7]. The results showed a correlation of 0.94 for the generated model compare with 0.91 from *Bjellqvist* function and standard deviation of 0.37 relative to 0.44 (Supplementary Information).

One of the major drawbacks of previously reported algorithms is their poor performance in the basic region. This relates to the number of peptide identifications in this region [1,2,7]. The use of the SVM algorithm with multiple steps of cross-validation and feature selection improved the *pI* estimation dramatically, particularly in the basic pH range. When the SVM based algorithm was used (Fig. 2A), even those wells containing a low number of identified peptides (at 9.15, 8.85, 8.55, and 8.25 pH units, respectively) presented a low standard deviation (0.37, 0.13, 0.27, and 0.12, respectively) in the theoretical *pI* values. The highest standard deviation (0.80) was observed in fraction 14 (at 7.6 pH units), where only 6 peptides were identified.

The use of isoelectric point as an orthogonal variable to support protein and peptide identification has been study recently [7,8,24,25]. Cargile and cols reported the theoretical basis for a new paradigm for identification. This methodology employs the use of accurate mass and peptide isoelectric point (*pI*) as identification criteria, and represents a change in focus from current tandem mass spectrometry-dominated approaches [26]. Also our group has previously reported the possibility of identifying theoretically peptides and proteins based on different experimental properties [8]. However, the use of isoelectric point as complement information to reduce the number of false positive peptide identifications hasn't been extensively exploited so far.

Table 2 shows the relation between the isoelectric point prediction *vs* the PeptideProphet probability. The isoelectric point range for a fraction is defined as the mean of predicted *pI* of the fraction ±2 standard deviation (stdev). A previous study demonstrated for different search engines that ±2 *pI* stdev had a stronger effect than ±1 *pI* stdev [7]. Our results show a low number of peptides (0.2%) with the highest PeptideProphet probabilities (1.0) fall outside the predicted range. The opposite effect was found for the peptides with the lowest PeptideProphet probabilities. This means that the isoelectric point prediction method can detect the number of possible false positive identifications for each fraction. Heller and cols [7] suggested in a previous study that for identifications

**Table 2 – The relation between (a) PeptideProphet probability and the (d) percentage of peptide with isoelectric point falls outside the predicted range (mean of the fraction ±2 standard deviation) for SVM Model. Non-redundant identifications row (e) is the number in almost one fraction that falls outside the predicted range. Column (b) is the number of identified peptides in each probability and (c) the number of non-redundant peptides in each probability. The total number of peptides outside the predicted *pI* was 750 peptides.**

| Probability[a] | 1 | 0.9 | 0.8 | 0.7 | 0.6 | 0.5 | 0.4 | 0.3 | 0.2 | 0.1 |
|---|---|---|---|---|---|---|---|---|---|---|
| Identified peptides[b] | 211687 | 33492 | 15960 | 11244 | 9780 | 9540 | 10200 | 11556 | 16212 | 4344 |
| Non-redundant peptides[c] | 16893 | 2791 | 1330 | 937 | 815 | 795 | 850 | 963 | 1351 | 362 |
| % peptides[d] | 0.2 | 2.6 | 5.9 | 6.1 | 9.3 | 14.0 | 16.4 | 16.8 | 22.6 | 31.2 |
| Non-redundant[e] | 10 | 34 | 39 | 33 | 45 | 68 | 94 | 113 | 228 | 86 |

with high PeptideProphet confidence scores, there are 2.9% false positives when applied the *Bjellqvist* function as filter. But, when they accounted the retention time additionally, they found more than 8.4% false positives [7]. When we applied the SVM approach to *Heller* data, we can detect more than 4.1% false positives for the high confidence identifications, which they cannot detect with its isoelectric point prediction alone.

Considering potential errors that can arise in the electrophoresis experimental protocol (focalization time, peptide abundance, peptide-peptide interactions, and sample composition, among others). The accuracy of the model allows us to find 44 non-redundant peptide identifications as likely false positives, all of which have high PeptideProphet probabilities (probabilities: 1, 0.9). Also for low identification probabilities (probabilities: 0.8–0.1), the algorithm identified more than 700 peptides in the experiment with theoretical isoelectric point outside the range of its fraction. Therefore, the method described in this manuscript could be used to rank the peptide identifications using orthogonal information, as suggested by previous studies [8,26].

In conclusion, we combined a SVM approach with only two simple peptide descriptors to predict the isoelectric point of identified peptides, and our results have shown better accuracy than the existing methods. Furthermore, the ability of calculating the *pI* of peptides to this accurate level is desirable for peptide *pI* filtering. We envisage that the same approach could also be applied to predict the effect of posttranslational modifications. The use of SVMs and the approach described in this work could be useful for these types of analyses.

Supplementary materials related to this article can be found online at doi:10.1016/j.jprot.2012.01.029.

## Acknowledgements

The authors would like to thank the INSPUR Company from China for its kind donation of the computer cluster TS10000 used for all calculations with the tools developed in this manuscript. JAV is supported by the EU FP7 grants LipidomicNet [grant number 202272] and ProteomeXchange [grant number 260558].

## REFERENCES

[1] Bjellqvist B, Hughes GJ, Pasquali C, Paquet N, Ravier F, Sanchez JC, et al. The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. Electrophoresis 1993;14:1023–31.

[2] Cargile BJ, Sevinsky JR, Essader AS, Eu JP, Stephenson Jr JL. Calculation of the isoelectric point of tryptic peptides in the pH 3.5–4.5 range based on adjacent amino acid effects. Electrophoresis 2008;29:2768–78.

[3] Vapnik V. The nature of statistical learning theory. Springer-Verlag New York, Inc.; 1995

[4] Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 2008;36:D202–5.

[5] Righetti PG. Determination of the isoelectric point of proteins by capillary isoelectric focusing. J Chromatogr A 2004;1037:491–9.

[6] Gauci S, van Breukelen B, Lemeer SM, Krijgsveld J, Heck AJ. A versatile peptide pI calculator for phosphorylated and N-terminal acetylated peptides experimentally tested using peptide isoelectric focusing. Proteomics 2008;8:4898–906.

[7] Heller M, Ye M, Michel PE, Morier P, Stalder D, Junger MA, et al. Added value for tandem mass spectrometry shotgun proteomics data validation through isoelectric focusing of peptides. J Proteome Res 2005;4:2273–82.

[8] Perez-Riverol Y, Sanchez A, Ramos Y, Schmidt A, Muller M, Betancourt L, et al. In silico analysis of accurate proteomics, complemented by selective isolation of peptides. J Proteomics 2011;74:2071–82.

[9] Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 2003;31:3784–8.

[10] Liu HX, Zhang RS, Yao XJ, Liu MC, Hu ZD, Fan BT. Prediction of the isoelectric point of an amino acid based on GA-PLS and SVMs. J Chem Inf Comput Sci 2004;44:161–7.

[11] Tian F, Yang L, Lv F, Zhou P. Predicting liquid chromatographic retention times of peptides from the *Drosophila melanogaster* proteome by machine learning approaches. Anal Chim Acta 2009;644:10–6.

[12] Supek F, Peharec P, Krsnik-Rasol M, Smuc T. Enhanced analytical power of SDS-PAGE using machine learning algorithms. Proteomics 2008;8:28–31.

[13] Lo SL, Cai CZ, Chen YZ, Chung MC. Effect of training datasets on support vector machine prediction of protein–protein interactions. Proteomics 2005;5:876–84.

[14] Burges CJC. A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 1998;2:121–67.

[15] Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics 2004;20:1466–7.

[16] Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 2002;74:5383–92.

[17] Kuhn M. Building predictive models in R using the caret package. J Stat Softw 2008;28:1–26.

[18] Karatzoglou A, Feinerer I. Kernel-based machine learning for fast text mining in {R}. Comput Stat Data Anal 2009.

[19] Zimmerman JM, Eliezer N, Simha R. The characterization of amino acid sequences in proteins by statistical methods. J Theor Biol 1968;21:170–201.

[20] Yang D, Ramkissoon K, Hamlett E, Giddings MC. High-accuracy peptide mass fingerprinting using peak intensity data with machine learning. J Proteome Res 2008;7:62–9.
[21] Timm W, Scherbart A, Bocker S, Kohlbacher O, Nattkemper TW. Peak intensity prediction in MALDI-TOF mass spectrometry: a machine learning study to support quantitative proteomics. BMC Bioinformatics 2008;9:443.
[22] Larranaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, et al. Machine learning in bioinformatics. Brief Bioinform 2006;7:86–112.
[23] Barla A, Jurman G, Riccadonna S, Merler S, Chierici M, Furlanello C. Machine learning methods for predictive proteomics. Brief Bioinform 2008;9:119–28.
[24] Horth P, Miller CA, Preckel T, Wenz C. Efficient fractionation and improved protein identification by peptide OFFGEL electrophoresis. Mol Cell Proteomics 2006;5:1968–74.
[25] Cargile BJ, Talley DL, Stephenson Jr JL. Immobilized pH gradients as a first dimension in shotgun proteomics and analysis of the accuracy of pI predictability of peptides. Electrophoresis 2004;25:936–45.
[26] Cargile BJ, Stephenson Jr JL. An alternative to tandem mass spectrometry: isoelectric point and accurate mass for the identification of peptides. Anal Chem 2004;76:267–75.