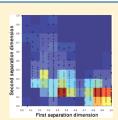


Orthogonality of Two-Dimensional Separations Based on Conditional Entropy

Mohammad Reza Pourhaghighi, Mohammad Karzand, and Hubert H. Girault*,

ABSTRACT: A new approach to assess the orthogonality of two-dimensional (2-D) separation systems based on conditional entropy is developed. It considers the quantitative distribution of peaks in the entire separation space such that the orthogonality obtained is independent of the number of peaks observed for each separation technique. Therefore, it can be used to compare the orthogonality of different 2-D separation protocols for a given sample. Herein, the developed method has been employed to estimate the orthogonality of peptide separation by off-gel electrophoresis (OGE) hyphenated to capillary zone electrophoresis (CZE).



The separation of complex samples requires the hyphenation of different separation techniques, as a single separation method does not usually possess a peak capacity sufficient to separate all the components. In 2-D separation systems, the theoretical peak capacity $(P_{\rm 2D})$ can be calculated by multiplying the individual peak capacities of each dimension. Nevertheless, the theoretical peak capacity can barely be attained and is often limited to the practical peak capacity $(N_{\rm p})$. As explained by Giddings, to achieve the maximum peak capacity $(P_{\rm 2D})$, the two separation mechanisms must be completely independent and the peaks must uniformly occupy the 2-D separation space. Since the evaluation of the orthogonality degree is essential to estimate the resolving power of different multidimensional separation protocols, it is important to develop a method that can cope with a wide range of experimental data.

Different approaches have already been developed to evaluate the orthogonality in 2-D separation systems. Liu et al. developed a geometrical approach based on the factor analysis.³ They used the retention times and capacity factors of each separation dimension to establish a correlation matrix and a peak spreading angle matrix. The orthogonality is then defined by a correlation matrix with correlation coefficients that vary from 0 for an ideally orthogonal to 1 for a nonorthogonal system. The main drawback of this approach is that the calculation of the orthogonality is based on the geometric distribution of the peaks along the diagonal of the 2-D separation space, which is not enough to describe the orthogonality between two separation methods, especially when the analytes are not diagonally distributed in the 2-D separation space.

Through a comprehensive study of different liquid chromatography (LC) modes for peptide separation, Gilar et al. proposed a simple geometrical approach to evaluate the orthogonality of different 2-D separation protocols. In their approach, a normalized 2-D separation space is first plotted and data points are placed into rectangular bins. With the total number of bins ($P_{\rm max}$) and the number of bins occupied by data points (Σ bins) known, the

orthogonality of a 2-D separation system is calculated by the following equation:

$$O = \frac{\sum \text{bins} - \sqrt{P_{\text{max}}}}{0.63P_{\text{max}}} \tag{1}$$

In Gilar's approach, the surface coverage of a normalized separation space varies from 10% in a nonorthogonal system to 63% for an ideally orthogonal system. Since 0.63 is only valid for some P values (= $(P_{\text{max}})^{1/2}$), Watson et al. modified eq 1 as follows:⁵

$$O = \frac{\sum \text{bins} - P}{0.63P^2 - P} \tag{2}$$

The main advantage of Gilar's geometrical approach was its simplicity, but since the orthogonality calculated only depends on the separation space surface coverage and not on the distribution pattern of the peaks, it does not describe the correlation between two separation mechanisms. For instance, if 50 out of 100 bins in a separation space are occupied by peaks (50% surface coverage), the orthogonality calculated using eq 2 is 75% without considering how these 50 bins are distributed. Furthermore, the surface coverage depends on the number of data points in the separation space, which undermines the orthogonality comparison when using different data sets.

Slonecker et al. proposed to describe the orthogonality of 2-D separation systems by information theory. Therein, "informational similarity" was used to describe the orthogonality of the 2-D separation system and its value varied between 0 for a completely orthogonal and 1 for a nonorthogonal 2-D separation system. Moreover, the % synentropy that was determined by dividing the informational entropy from data diagonally aligned by the total 2-D informational entropy was introduced to describe the degree

Received: April 5, 2011
Accepted: September 13, 2011
Published: September 13, 2011

[†]Laboratoire d'Electrochimie Physique et Analytique, Station 6, Ecole Polytechnique Fédérale de Lausanne, CH-1015, Lausanne, Switzerland

[‡]Information Theory Laboratory, Station 14, Ecole Polytechnique Fédérale de Lausanne, CH-1015, Lausanne, Switzerland

of nonorthogonality along the diagonal of the 2-D separation space. Consequently, a synentropy percentage equal to 0% describes a 2-D separation system in which the two dimensions are completely orthogonal. Despite a valuable effort to expand the application of information theory to orthogonality calculations by using different descriptors to estimate the orthogonality of the separation systems, the method proposed is not able to describe the orthogonality of 2-D systems where the correlation between the separation mechanisms is not along the diagonal (i.e., off-diagonal correlations).

In the present study, we propose a novel approach to evaluate the orthogonality of 2-D separation systems based on conditional entropy. This approach considers the quantitative data distribution in the entire separation space, as off-diagonal correlations between the two separations mechanisms are also considered in the orthogonality calculation. Furthermore, since the orthogonality calculated is independent of the number of peaks in the separation space, a methodology based on conditional entropy can be employed to compare, for a given sample, the orthogonality of different 2-D separation protocols. For a peptide separation, it is for example possible to compare the merits of hyphenating different protocols such as reverse phase chromatography coupled to strong cation exchange chromatography (RP/SCX) and off-gel electrophoresis to capillary zone electrophoresis (OGE/CZE).

■ EXPERIMENTAL SECTION

Materials and Reagents. All chemicals used were of analytical grade and obtained from Sigma-Aldrich (Schnelldorf, Switzerland). All buffer and protein solutions were prepared with water produced by an alpha Q Millipore system (Zug, Switzerland).

Tryptic Digest. Bovine serum albumin (BSA), myoglobin (Myo), β -lactoglobulin (β -Lac) and cytochrome C (Cyt. C) were dissolved in 50 mM ammonium bicarbonate (pH 8.2) and heated at 100 °C for 5 min. Then, trypsin was added to the protein solution with a 1:100 enzyme to protein ratio, and the tryptic digestion was performed overnight at 37 °C.

Off-Gel Electrophoresis. OGE separations were performed with the Agilent 3100 OFFGEL fractionator (Waldbronn, Germany). An 18 cm immobilized pH gradient (IPG) strip pH 3-10 (Amersham Biosciences, Otelfingen, Switzerland) was used for the experiments allowing the collection of 18 fractions. The focusing was carried out with voltage and current limited to 4.5 kV and 150 μ A, respectively, and stopped after achieving 45 kVh. At the end of the fractionation, the peptide solution in each well was collected and further analyzed by CE without any particular treatment.

Capillary Electrophoresis. CE experiments were performed with a Hewlett-Packard^{3D} CE system (Waldbronn, Germany). Fused silica capillaries (50 μm i.d., 26.5 cm effective length, 35 cm total length) were obtained from BGB Analytik AG (Boeckten, Switzerland) and coated with 5% hydroxypropyl cellulose (HPC) in the laboratory according to the procedure described by Shen et al.⁷ Phosphate buffer (66 mM) pH 3.0 was used as background electrolyte (BGE) for CZE separations. Samples were injected electrokinetically (2 kV, 60 s), and peptide separation was performed by applying 20 kV across the capillary (0.57 kV/cm) while the UV absorbance of analytes was monitored at 200 nm. Precise analysis of the electropherograms was performed with 32 Karat software (Beckman Coulter, CA), and migration times of the peaks detected were first normalized and

then transferred to a program written in MATLAB for orthogonality calculation.

Orthogonality Calculation. A program was written in MATLAB⁸ to calculate the information entropy for each separation dimension, the joint entropy, the mutual information (in bits), as well as the orthogonality of the 2-D separation system as are explained in the Theory section.

■ THEORY

Information Entropy. Information theory was first developed by Shannon. In this theory, entropy is defined as a measure of the random variable uncertainty. Suppose X is a discrete random variable within an alphabet χ , then the probability mass function $p(x) = \Pr$, $x \in \chi$, the entropy of X is defined by (in unit of bits):

$$H(X) = -\sum_{x \in \gamma} p(x) \log_2 p(x)$$
 (3)

Note that the entropy is a function of the distribution of X and does not depend on the actual value of X but only on the probability of occurrence of each output.

Similarly, the joint entropy of a pair of discrete random variables, H(X,Y), with a joint distribution p(x,y) can be defined as

$$H(X,Y) = -\sum_{x \in \chi} \sum_{y \in Y} p(x,y) \log_2 p(x,y)$$
 (4)

As already described in refs 6 and 11 in a 2-D separation, p(X) is the probability of a peak to appear at a particular retention time. Accordingly, referring to the distribution of peaks at different retention times, the entropy of each individual separation dimension as well as the joint entropy of the entire 2-D separation system can be calculated using eqs 3 and 4, respectively.

Conditional Entropy. The entropy of the variable Y conditioned on the variable X quantifies the remaining entropy or uncertainty of a random variable Y, conditioned on the variable X taking a certain value x and is written as H(Y|X) and defined as X

$$H(Y|X) = \sum_{x \in \chi} p(x)H(Y|X = x)$$
 (5)

From these definitions, the entropy of Y conditional on X can be obtained from the following equation (chain rule for conditional entropy):

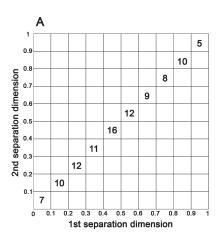
$$H(X,Y) = H(Y|X) + H(X) \tag{6}$$

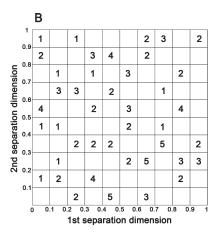
Considering X and Y as the two dimensions of a 2-D separation protocol, the H(X,Y) bit of information is needed to reconstruct the 2-D system. With the values of the first dimension (X) revealed (e.g., the retention times), H(X) bits of information are known and H(Y|X) bits of uncertainty are still remaining in the 2-D system.

Consequently, if and only if the 2-D separation system is completely nonorthogonal, for instance the retention times in the second dimension (Y) are completely determined by the first dimension (X), $H(Y \mid X)$ is equal to zero. On the contrary, in a fully orthogonal system where the two separation mechanisms (X and Y) are completely independent, $H(Y \mid X)$ is equal to H(Y).

Therefore, we propose the following equation to quantify the orthogonality in a 2-D separation system

$$O\% = \frac{H(Y|X)}{H(Y)} \times 100 \tag{7}$$





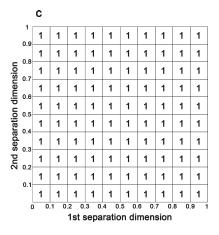


Figure 1. Orthogonality calculation based on conditional entropy. The numbers represent the number of peaks in each individual division of the normalized separation space: (A) nonorthogonal system, O = 0%; (B) random distribution, partially orthogonal system, O = 59%; (C) fully (ideal) orthogonal system, O = 100%.

where H(Y|X) represents the entropy of the second dimension conditional on the first dimension and H(Y) is the entropy of the second dimension. The orthogonality obtained by this method varies between 0 for a nonorthogonal system and 100% for a fully orthogonal 2-D separation system.

■ RESULTS AND DISCUSSION

Figure 1 illustrates the principle of the present approach for orthogonality evaluation of three different hypothetical 2-D separation spaces. To evaluate the orthogonality in 2-D separation systems, a square matrix that represents the normalized 2-D separation space must initially be reconstructed. Normalization of the retention times, $(R_t)_{ij}$, in each separation dimension was performed according to eq 8 where $(R_t)_{min}$ and $(R_t)_{max}$ represent the minimum and maximum retention times in all data sets, respectively.

$$(R_{\rm t})_{\rm norm} = \frac{(R_{\rm t})_i - (R_{\rm t})_{\rm min}}{(R_{\rm t})_{\rm max} - (R_{\rm t})_{\rm min}}$$
 (8)

Subsequently, considering the distribution of peaks in each separation dimension and in the entire 2-D separation space, the information entropy of each dimension as well as the information entropy of the entire 2-D separation space were calculated using eqs 3 and 4, respectively, with p(X) being the probability for the presence of a peak at a specific retention time. For example, the information entropy for the first separation dimension of the example shown in Figure 1B would be computed as

$$H(X) = -\left[\left(\frac{9}{100}\log_2\frac{9}{100}\right) + \left(\frac{8}{100}\log_2\frac{8}{100}\right) + \dots + \left(\frac{7}{100}\log_2\frac{7}{100}\right)\right] = 3.3 \text{ bits}$$
 (9)

Table 1 presents the values obtained for the information entropy of each separation dimension and the entire 2-D separation space, the entropy of the second separation dimension conditional on the first dimension, as well as the orthogonality obtained for the three examples shown in Figure 1.

In Figure 1A, all the peaks are positioned on the diagonal of the separation space. This example represents a 2-D separation

Table 1. Different Information Theory Based Parameters Calculated for the 2-D Separations Described in Figure 1^a

	H(X)	H(Y)	H(X,Y)	H(Y X)	O %
1-A	3.26	3.26	3.26	0	0
1-B	3.30	3.26	5.23	1.93	59
1-C	3 32	3 32	6.64	3.32	100

 $^aH(X)$, informational entropy of first separation dimension in bits; H(Y), informational entropy of second separation dimension in bits; H(X,Y), Jjoint entropy of entire 2-D separation system in bits; H(X|Y), entropy of second separation dimension conditional on first separation dimension in bits; O(Y), orthogonality degree in percent.

with identical separation mechanisms in both dimensions. Since the data distribution in both dimensions is identical, the entropy of both separation dimensions and joint entropy are equal, which indicates the maximum correlation. As expected, H(Y|X) and the orthogonality are equal to zero. The orthogonality of any similar situation where the normalized matrix contains only one nonzero element in each column would also be zero. Indeed, the zero orthogonality of these examples could be explained by the fact that no improvement in separation has been achieved by employing the second separation dimension.

In Figure 1B, the peaks are randomly distributed in the 2-D separation space. This is the situation mostly encountered in practice where only some of the bins in the separation space are used. The 59% orthogonality demonstrates a partially orthogonal separation. Finally, Figure 1C shows an ideal case where the separation space is uniformly covered by data points. In this case, the entropies for both separation dimensions are equal, and since H(Y|X) = H(Y), the two separation techniques are completely independent and 100% orthogonality is achieved.

Bin Number. As described above, in order to estimate the entropy of each separation dimension, the probability distributions have to be calculated. Herein, histograms are used to estimate the probability data distribution. The number of bins in the histogram is determined by the data range and the bin width. Choosing a very small bin width will result in many bins, and the frequency distribution will look like a broken comb, which does not really represent a real data distribution. On the contrary, setting the bin width to an excessively large value will result in a small number of bins and the distribution contains too

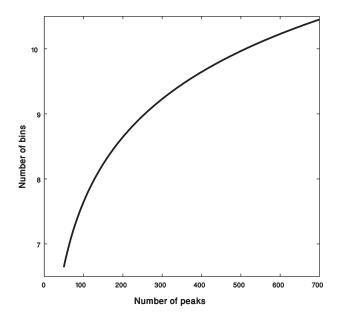


Figure 2. Optimum bin number in each separation dimension depending on the total number of peaks in separation space.

little information to be useful. The only general rule is that the ideal number of bins is related to the size of the data set. If the frequency distribution tabulates the frequency of a huge number of values, it makes sense to use a small bin width. If the frequency distribution is for a small data set, a larger bin width makes sense. Many algorithms were devised to define the ideal bin width. Here, on the basis of Sturges' method, ¹² the following equation was employed to optimize the number of bins.

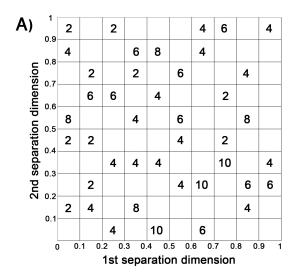
$$bin count = 1 + log_2(peak numbers)$$
 (10)

$$bin width = \frac{[(R_t)_{max} - (R_t)_{min}]}{bin count}$$
(11)

The optimum number of bins in each separation dimension versus the number of peak present in the 2-D separation space is presented in Figure 2. Although varying the bin number may cause minute changes in the orthogonality calculated (normally less than 5%), considering that the number of peaks in most 2-D separations varies between 100 and 500, the separation space was here divided into 10×10 rectangular bins to illustrate the method for a peptide separation.

Effect of the Peak Number. Since each individual sample used in 2-D separation provides a different number of peaks, it is important that the orthogonality assessment methodology is independent of the total number of data points in the separation space. As explained before, the entropy of each separation dimension is a function of the peak distribution along the respective separation axis and does not depend on the number of the peaks. Therefore, except the abnormal cases with very low number of peaks (i.e., 1, 2), the orthogonality calculated by this conditional entropy method is independent of the total number of peaks in the separation space. This feature is a key advantage when comparing the orthogonality of different 2-D separation methodologies.

Figure 3 illustrates the effect of the peak number and their distribution in the separation space on the calculated orthogonality. The two examples shown in Figure 3 have a geometrical distribution similar to Figure 1B. Figure 3A shows an example where the number of peaks in the separation space is uniformly



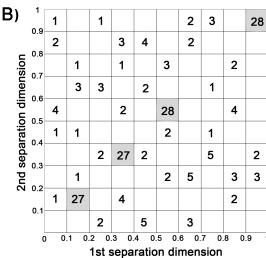
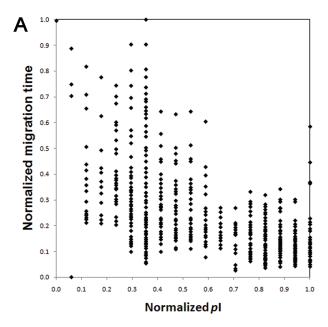


Figure 3. Effect of increasing the number of peaks on orthogonality calculated with hypothetical examples: (A) O = 60% and (B) O = 43%.

doubled relative to Figure 1B without disturbing the geometrical distribution pattern. In spite of twice the number of peaks in the separation space, the orthogonality of the 2-D system remains unchanged (59%) since there is no evidence to demonstrate higher or lower correlation between the two separation mechanisms. In Figure 3B, doubling the total number of peaks by increasing the number of peaks on the diagonal of the normalized separation space only points toward more correlation between the two separation mechanisms. As a consequence, the orthogonality of such a system based on conditional entropy is reduced to 43%.

These models demonstrate that even if the method proposed is sensitive to quantitative peak distribution in the separation space, it does not depend on the number of peaks except if it changes the correlation between the two separation mechanisms.

Orthogonality of OGE-CZE. Off-gel electrophoresis is a technique developed in our laboratory for high-resolution fractionation of peptides and proteins according to their isoelectric point (pI) at the micropreparative scale. ^{13,14} Taking advantage of the reproducibility of the IPGs, the separated compounds are recovered in solution unlike in classical gel isoelectric focusing (IEF). In comparison to capillary IEF (CIEF), a lower concentration of carrier ampholytes (CAs) can be used and neither anolyte



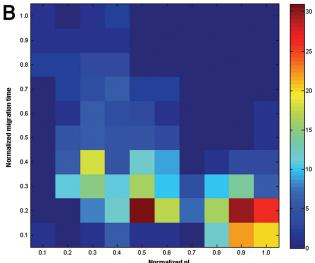


Figure 4. Normalized 2-D OGE-CZE separation plot of standard peptide mixture (separation conditions are described in the Experimental Section): (A) actual peak distribution and (B) graphic illustration of quantitative peak distribution in normalized separation space used for orthogonality calculation.

nor catholyte are needed for separation. These features facilitate the integration of OGE into any proteomics workflow as demonstrated by its previous use as a first dimension before liquid chromatography—tandem mass spectrometry (LC—MS/MS)^{15–18} and CZE.¹⁹

Busnel et al. showed that the hyphenation of OGE with CZE can be successfully employed for the high-resolution separation of complex peptide samples. ¹⁹ Furthermore, the orthogonality of the OGE-CZE hyphenation estimated by Gilar's geometrical method was determined to be comparable with 2-D LC separation systems.

Herein, to demonstrate the potential of the present approach for orthogonality evaluation, a standard peptide mixture containing tryptic digest of BSA, myoglobin, Cyt. C, and β -Lac was first separated by OGE and then each off-gel fraction was analyzed by CZE as the second separation dimension. The normalized 2-D

separation plot obtained is shown in Figure 4A. Because of the acidic BGE (pH 3.0) used for the CZE separation of the off-gel fractions, the number of peaks in the acidic pIs of the separation space is less than expected. Consequently, the best separation is obtained for off-gel fractions with pIs ranging from 4 to 7. For more basic off-gel fractions, as shown in figure 4, the separation efficiency is limited by a reduced migration time of highly charged peptides.

Afterward, to calculate the orthogonality of the separation, since about 440 peaks are present in this separation space, the normalized separation space was divided to 10×10 rectangular bins. Figure 4B demonstrates the quantitative distribution of peaks in divided normalized 2-D separation space. Accordingly, the informational entropy of each separation dimension, the joint entropy of the entire 2-D separation system, and the entropy of the second dimension (CZE) conditional to the first dimension (OGE) were calculated using the aforementioned equations. Finally, the orthogonality of CZE-OGE hyphenation with the conditional entropy approach was calculated by eq 7 to be 86%.

As a comparison, the highest degree of orthogonality for peptide separation by 2D-LC systems that is reported refers to the hyphenation of hydrophilic interaction chromatography and reverse-phase chromatography (HILIC-RP 2D-LC)⁴ and was also calculated by a conditional entropy approach with a value of about 86%.

It is important to notice that while the orthogonality of a 2-D separation describes the potential correlation between two dimensions, other parameters such as the 2-D peak capacity are also required to evaluate the separation efficiency. Therefore, the practical peak capacity of OGE-CZE hyphenation was also calculated using the following equation:

$$N_{\rm p} = (P_1 P_2) O \tag{12}$$

 $N_{\rm p}$ represents the practical peak capacity, P_1 and P_2 are the respective peak capacities of the first and second dimensions, and O is the orthogonality degree of the 2-D separation system. In the 2-D OGE-CZE experiment presented here, the peak capacity of OGE corresponds to the number of fractions, which is equal to 18. In the second dimension, CZE separation of each OGE fraction, the average peak width was calculated to be about 0.035 min over a separation window of 11.3 min providing a peak capacity around 323. As a result, from eq 8, the practical peak capacity of 2-D OGE-CZE separation of peptides is about 5800.

■ CONCLUSIONS

A novel approach to evaluate the orthogonality in 2-D separation systems based on conditional entropy is presented. Compared with previous methods for orthogonality calculation, the present approach considers the quantitative peak distribution in the entire 2-D separation space. Therefore, even off-diagonal correlations are considered. Moreover, since the orthogonality calculated by conditional entropy does not depend on the number of peaks in the separation space but on their quantitative distribution, the method developed can be employed to compare the orthogonality of different 2-D separation protocols.

AUTHOR INFORMATION

Corresponding Author

*E-mail: hubert.girault@epfl.ch. Phone: [+41 21 69] 33151. Fax: [+41 21 69] 33667.

■ ACKNOWLEDGMENT

The authors would like to thank "Agilent Technologies Foundation" for financial support.

■ REFERENCES

- (1) Giddings, J. C. J. High Resolut. Chromatogr. 1987, 10, 319-323.
- (2) Giddings, J. C. Unified Separation Science; Wiley: New York, 1991.
- (3) Liu, Z.; Patterson, D. G.; Lee, M. L. Anal. Chem. 1995, 67, 3840–3845.
- (4) Gilar, M.; Olivova, P.; Daly, A. E.; Gebler, J. C. Anal. Chem. 2005, 77, 6426–6434.
- (5) Watson, N. E.; Davis, J. M.; Synovec, R. E. Anal. Chem. 2007, 79, 7924–7927.
- (6) Slonecker, P. J.; Li, X.; Ridgway, T. H.; Dorsey, J. G. Anal. Chem. 1996, 68, 682–689.
 - (7) Shen, Y.; Smith, R. D. J. Microcolumn Sep. 2000, 12, 135-141.
- (8) MATLAB, version 7.9. (R2009b); The Math Work Inc.: Natick, MA, 2009; http://www.mathworks.com.
 - (9) Shannon, C. E. Bell Syst. Tech. J. 1948, 27, 623-656.
- (10) Cover, T, M; Thomas, J. A. Elements of Information Theory; Wiley: New York, 2006.
 - (11) David, V.; Medvedovici, A. J. Chemom. 2005, 19, 16-22.
 - (12) Sturges, H. A. J. Am. Stat. Assoc. 1926, 21, 65-66.
- (13) Ros, A.; Faupel, M.; Mees, H.; Van Oostrum, J.; Ferrigno, R.; Michel, P.; Rossier, J. S.; Girault, H. H. *Proteomics* **2002**, *2*, 151–156.
- (14) Michel, P. E.; Reymond, F.; Arnaud, I. L.; Josserand, J.; Girault, H. H.; Rossier, J. S. *Electrophoresis* **2003**, *24*, 3–11.
- (15) Heller, M.; Michel, P. E.; Morier, P.; Crettaz, D.; Wenz, C.; Tissot, J. D.; Reymond, F.; Rossier, J. S. *Electrophoresis* **2005**, 26, 1174–1188.
- (16) Geiser, L.; Dayon, L.; Vaezzadeh, A. R.; Hochstrasser, D. F. *Methods Mol. Biol.* **2011**, 681, 459–472.
- (17) Waller, L. N.; Shores, K.; Knapp, D. R. J. Proteome Res. 2008, 7, 4577–4584.
- (18) Michel, P. E.; Crettaz, D.; Morier, P.; Heller, M.; Gallot, D.; Tissot, J. D.; Reymond, F.; Rossier, J. S. *Electrophoresis* **2006**, 27, 1169–1181.
- (19) Busnel, J. M.; Lion, N.; Girault, H. H. Anal. Chem. 2007, 79, 5949–5955.