

Chapter 2

Decoding 2-D Maps by Autocovariance Function

Maria Chiara Pietrogrande, Nicola Marchetti, and Francesco Dondi

Abstract

This chapter describes a mathematical approach based on the study of the 2-D autocovariance function (2-D ACVF) useful for decoding the complex signals resulting from the separation of protein mixtures. The method allows to obtain fundamental analytical information hidden in 2-D PAGE maps by spot overlapping, such as the number of proteins present in the sample and the mean standard deviation of the spots, describing the separation performance. In addition, it is possible to identify ordered patterns potentially present in spot positions, which can be related to the chemical composition of the protein mixture, such as post-translational modifications.

The procedure was validated on computer-simulated maps and successfully applied to reference maps obtained from literature sources.

Key words 2-D PAGE (2-D polyacrylamide gel electrophoresis) maps, Chemometric methods, Bidimensional autocovariance function, Spot overlapping, Bioinformatics

1 Introduction

Polyacrylamide gel electrophoresis (2-D PAGE) separation of proteins is considered the classical and principal tool for proteomic studies, combined with mass spectrometry, to achieve a comprehensive identification and quantification of almost every protein present in a complex biological (animal or plant tissue) sample [1, 2]. However, in the last 10 years the importance of two- and multi-dimensional separation techniques (mainly 2-D liquid chromatography) has been considerably increased, as witnessed by recent advances in shotgun or top-down approaches [3, 4].

Despite the enormous improvements in chemical technologies and separation efficiency [5–8], and some impressive innovations in robotics and algorithms for spot detection, protein databases, and data handling [9, 10], a comprehensive separation and elucidation of all the proteins present in the sample is still far from being achieved [11–14]. This is due to the intrinsic complexity of cells and biological fluids, that can contain thousands of proteins, present in a wide range of relative abundances and displaying great

differences in structure and size. The consequence is that co-migrating proteins can be easily present in the same spot, which results in a drop of the quality of the analytical information contained in the map [13, 14]. Therefore, plethora of data obtained from each analytical run require a proper signal processing procedure for decoding the complexity of the 2-D map, in order to fully extract the whole analytical information contained therein, in particular, information very relevant to proteomics, such as the number of proteins and the chemical composition of the sample, i.e., proteins occurrence, identity, abundance, and chemical structure.

The mathematical-statistical method herein described is based on the study of the 2-D autocovariance function (2-D ACVF), computed on the experimental digitized map, i.e., experimental 2-D ACVF (2-D EACVF) [15–18]. This method allows to estimate the complexity of the mixture (number of components, abundance distribution) and the separation performance. Moreover, the study of the 2-D EACVF plot allows to identify the potential presence of structured patterns of protein spots, which can be related to specific protein structural modifications [17].

1.1 Theory

The experimental map is typically acquired in a digitized form consisting of a gridded surface $N_x \times N_y$, where all the nodes are equally spaced. As an example, a computer-generated map representing experimental 2-D PAGE gels is reported in Fig. 1a: it contains 200 proteins with an elliptic spot size of $\sigma_x = 0.025 \text{ pI}$ and $\sigma_y = 0.0006 \log Mr$.

The 2-D EACVF is computed on the digitized map as:

$$C_{k,l} = \frac{1}{N_x N_y} \sum_{i=1}^{N_x-k} \sum_{j=1}^{N_y-l} (f_{i,j} - \bar{f})(f_{i+k,j+l} - \bar{f}) \quad (1)$$

$$k = -M_x, \dots, -1, 0, 1, \dots, M_x$$

$$l = -M_y, \dots, -1, 0, 1, \dots, M_y$$

where:

N_x and N_y are the total number of points of the digitized map along the two separation axes,

$f_{i,j}$ represents the map intensity at the point (i, j) ,

\bar{f} is the average intensity calculated over all the sampled points,

k and l are the lags between subsequent points in the map along the two separation axes over which 2-D EACVF is calculated,

M_x and M_y are the maximum number of lags used for 2-D EACVF calculation.

Each point used for computation can be converted into $\Delta x = \Delta \text{pI}$ and $\Delta y = \Delta \log Mr$ on the basis of the sampling interdistances between subsequent points along the X and Y axes.

2-D EACVF can be plotted vs. the interdistances along the two separation axes (ΔpI and $\Delta \log Mr$) to obtain a 2-D EACVF plot: Fig. 1b reports the contour plot of 2-D EACVF computed on the 2-D map shown in Fig. 1a. It is characterized by a main peak centered at the origin with a bidimensional Gaussian shape (enlarged detail in Fig. 1b). Besides the peak in the center, some minor fluctuations of the autocovariance function are observed.

Theoretical expressions for 2-D ACVF (2-D TACVF) were derived to express 2-D TACVF as a function of the separation parameters. The theoretical model requires a frequency function for describing the distribution of spot positions over the X, Y area. Two limiting cases are considered, describing: 1) a completely disordered separation and 2) an ordered map [15]. In both cases the single protein spots are represented by a bivariate Gaussian distribution described by the standard deviations along the two separation axes, σ_x and σ_y ; both circular ($\sigma_x = \sigma_y$) and elliptical ($\sigma_x \neq \sigma_y$) spots are assumed (see **Note 1**).

1. *Completely disordered separation* is characterized by protein positions randomly distributed on the 2-D space (Poissonian retention pattern) (see **Note 2**): in this case the 2-D TACVF is given by the following equation [15]:

$$C(\Delta x, \Delta y) = \frac{V_T^2 (\sigma_h^2 / a_h^2 + 1)}{4\sigma_x \sigma_y \pi X Y m} e^{-[(\Delta x)^2 / 4\sigma_x^2 - (\Delta y)^2 / 4\sigma_y^2]} \quad (2)$$

where:

$V_T = 2\pi\sigma_x\sigma_y m a_h$ is the total volume of the signal computed on the three coordinates (x, y, f) ,

m is the number of detectable components,

X and Y are the lengths of the separation space,

σ_x and σ_y are the standard deviations of single protein spots along the two separation axes, and

σ_h^2 / a_h^2 describes the distribution of spot intensities, where a_h is the mean value of protein abundance and σ_h^2 its variance.

2. An *ordered pattern* in a 2-D PAGE map may be formed by ordered sequences of protein spots, where the position of the n -th term of the series is described by:

$$x(n) = a_x + b_x n \quad (3a)$$

$$y(n) = a_y + b_y n \quad (3b)$$

where a_x, a_y, b_x , and b_y are constants (see **Note 3**).

In this case, the expression of 2-D TACVF is [15]:

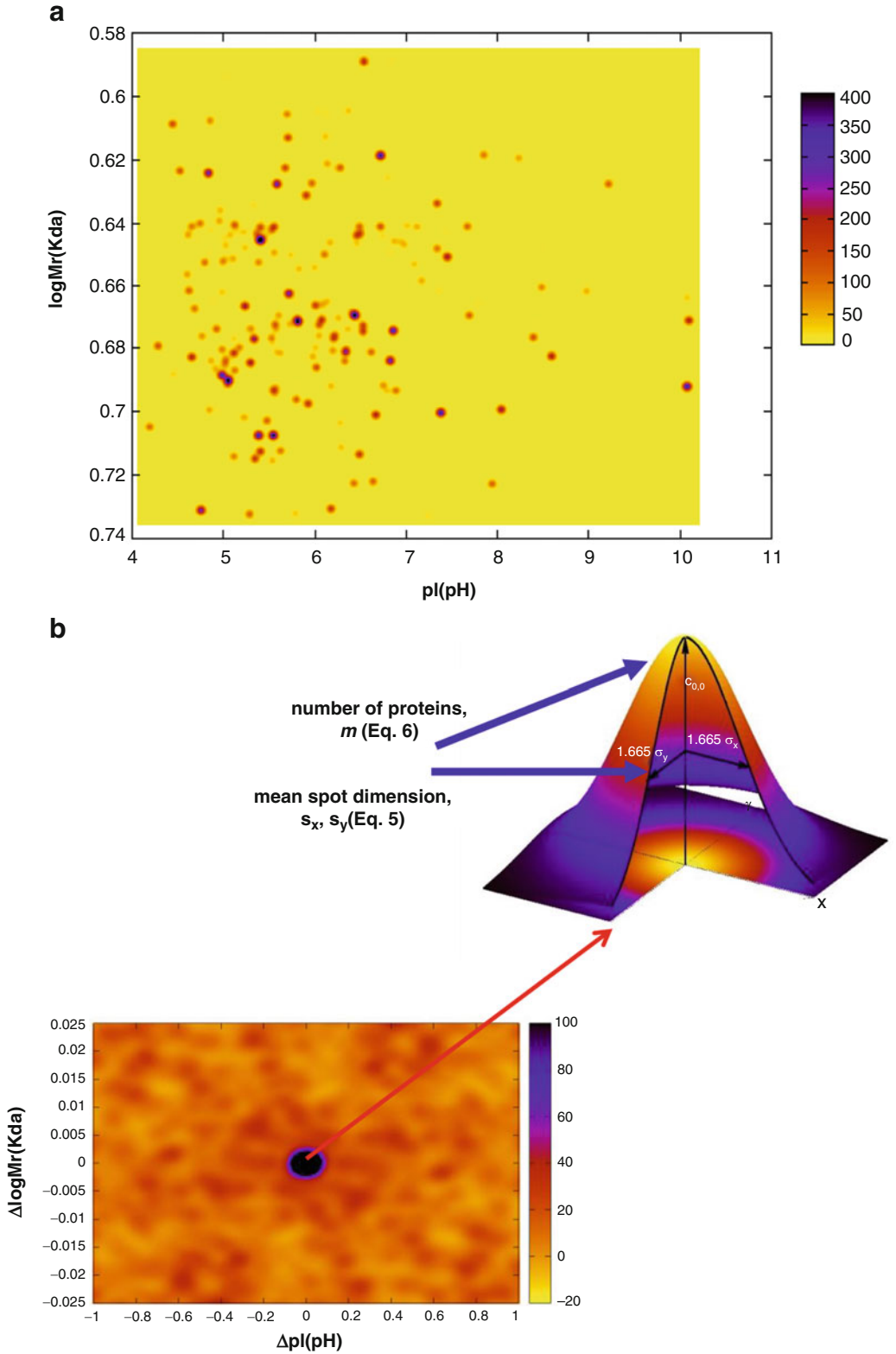


Fig. 1 Computation of the experimental 2-D EACVF on a 2-D generated map. **(a)** Simulated map containing 200 proteins and spot dimensions $\sigma_x = 0.026$ pI and $\sigma_y = 0.0006$ logMr. **(b)** 2-D EACVF plot versus interdistance along the two separation axes ΔpI and $\Delta \log Mr$; enlarged detail: 3-D view of the 2-D EACVF plot region for short

$$C(\Delta x, \Delta y) = \sum_{k=0}^{k=n_{\max}} \frac{V_T^2}{4\sigma_x\sigma_y\pi XY(n_{\max} - k + 1)} \left(\frac{\sigma_h^2}{a_h^2} + 1 \right) \times e^{-[(\Delta x - b_X k)^2 / 4\sigma_x^2] - [(\Delta y - b_Y k)^2 / 4\sigma_y^2]} \quad (4)$$

where n_{\max} is the highest value of n , i.e., the number of proteins of the series.

In this case, the 2-D TACVF plot shows well-defined cones located at interdistances kb_x and kb_y ; they are called deterministic since they correspond to repeated interdistances among the terms of the series. Their height decreases with k , but their shape is independent on k .

2 Materials

2.1 PC-Generated SDS 2-D PAGE Maps

1. Take into consideration datasets of the pI and $\log Mr$ coordinates from the SWISS-2-D PAGE database [19].
2. Retrieve the pI and $\log Mr$ values of identified spots of real reference maps of human tissues from SWISS-2-D PAGE database [19].

2.2 Software

1. Use the dedicated software Melanie (Geneva Bioinformatics, GeneBio S.A. Geneva, Switzerland) to detect the spots of the digitized maps and to measure their volumes.
2. Write the numerical calculation algorithms in Fortran.

3 Method

3.1 Calculation of PC-Generated 2-D Maps

1. Generate 2-D maps with known separation properties in order to validate the method [16–18].
2. Describe each spot by two position coordinates (pI and $\log Mr$) and by a third coordinate $f_{i,j}$ representing the spot intensity.
3. Generate the pI and $\log Mr$ coordinates that follow the same position distribution present in real maps, by applying the rejection algorithm: the flowchart of the proposed algorithm is reported in Fig. 2a. Such a distribution representing real maps was evaluated from 1956 identified spots in reference maps of human tissues retrieved from the SWISS-2-D PAGE database [19].
4. The rejection algorithm is based on a simple but efficient method for generating random coordinates whose distribution

Fig. 1 (continued) for interdistances ($\Delta pI \leq 4\sigma_x$ and $\Delta \log Mr \leq 4\sigma_y$) from which the parameters m , σ_x , and σ_y are estimated. Reproduced from ref. [16] with kind permission of the WILEY-VCH Verlag GmbH

a

FUNCTION RM

DEFINE $p(x)$

DEFINE $f(x)$

CALCULATE A

FOR $i=1$ to N

 ran1 = random number in $[0;A]$

 CALCULATE x_0

 CALCULATE $f(x_0)$

 REPEAT

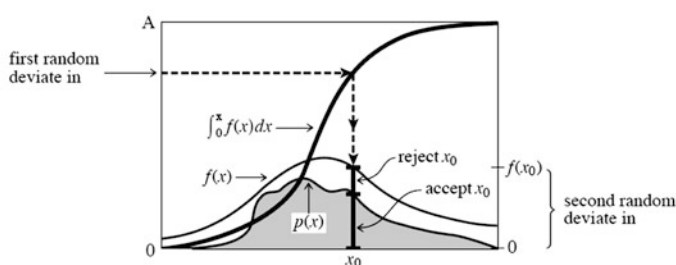
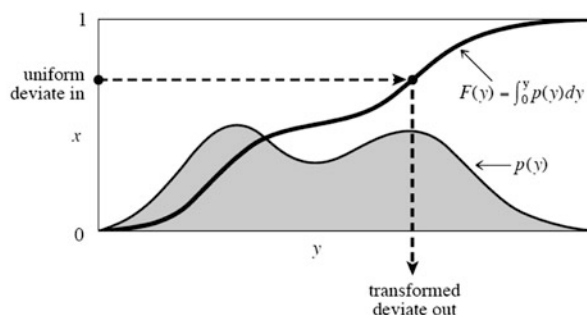
 ran2 = random number in $[0; f(x_0)]$

 UNTIL $\text{ran2} \leq p(x)$

$c1(i) = \text{ran2}$

NEXT i

END FUNCTION



b

FUNCTION 2D-EACVF

INPUT X (vector of x coordinates)

INPUT Y (vector of y coordinates)

INPUT Z (matrix of spot intensities)

fav = average value of elements $z(i,j)$

For $p=-M_x$ to M_x

 For $q=-M_y$ to M_y

 For $i=1$ to N_x-p

 For $j=1$ to N_y-q

$\text{EACVF}(p,q) = 1/(N_x \times N_y) \times \text{SUM}[(z(i,j) - \text{fav}) \times (z(i+p,j+q) - \text{fav})]$

 NEXT j

 NEXT i

 NEXT q

NEXT p

END FUNCTION

Fig. 2 Flowchart of the algorithms used for computation. (a) Rejection algorithm for generating the $p/$ and $\log Mr$ coordinates that follow the same position distribution present in real maps. (b) Calculation of the 2-D EACVF on the digitized map

function $p(x)$ is known within a given range of x but not analytically defined. The cumulative distribution function (the indefinite integral of the density probability function $p(x)$) is not required. Let's consider a probability distribution function $f(x)$ which has finite area and encloses $p(x)$ (*see* lower inset in Fig. 2a). $f(x)$ is called the comparison function. The integral of $f(x)$ has to be computable and A is the area below the curve. First, generate a random uniform deviate between 0 and A and pick the corresponding x_0 value from the cumulative distribution function of $f(x)$ (transformation method, upper inset in Fig. 2a). Then, independently generate a second random uniform deviate between 0 and $f(x_0)$. If this is lower than $p(x_0)$, then x_0 is accepted; otherwise, x_0 is rejected and the procedure repeated again from the beginning. More information on transformation and rejection methods can be found in dedicated textbooks [21].

5. In addition, build up reference 2-D maps using the pI and $\log Mr$ values of identified spots retrieved from real reference maps of human tissues (SWISS-2-DPAGE database) [19].
6. Generate spot intensity values using a random distribution that has been demonstrated the most probable for a high number of components [11, 20].
7. Describe spots with elliptical shape selecting proper σ_x and σ_y values to represent different experimental conditions, i.e., map dimension, pH gradient, and scanner resolution (*see* **Note 4**).

3.2 Numerical Calculation of 2-D EACVF on Digitized 2-D Maps

1. Calculate the 2-D EACVF on the digitized map, according to Eq. 1: the flowchart of the proposed algorithm is reported in Fig. 2b (*see* **Note 5**).
2. Plot it versus the interdistance along the two separation axes to obtain the 2-D EACVF plot: as an example, Fig. 1b shows the 2-D EACVF plot computed on the simulated map of Fig. 1a. In general, the 2-D EACVF plot clearly shows a maximum of 2-D EACVF at short interdistances, lower than the average spot sizes $4\sigma_x$ and $4\sigma_y$ (short-term correlations, shown in the inset in Fig. 1b), followed by lower values at higher interdistances (long-term correlations, $\Delta pI \geq 4\sigma_x$ and $\Delta \log Mr \geq 4\sigma_y$).

3.3 Estimation of the Separation Parameters

1. Here a simplified version of the 2-D ACVF approach is described: it is based on the measurement of fewer points than the experimental 2-D ACVF for estimating the separation parameters, i.e., number of detectable components (m) and the average spot widths ($4\sigma_x$ and $4\sigma_y$), and identifying the potential presence of ordered structures [16].

2. The first part of 2-D EACVF represents the mean spot size averaged on all the spots present in the map, described by $4\sigma_x$ and $4\sigma_y$. The pre-exponential term:

$$\frac{V_T^2(\sigma_h^2/a_h^2 + 1)}{4\sigma_x\sigma_y\pi XYm} \quad (5)$$

is the same in both random and ordered models (Eqs. 2 and 4).

3. Compute the spot mean standard deviation, $4\sigma_x$ and $4\sigma_y$, from the width of the 2-D EACVF bidimensional Gaussian peak at half height as (enlarged detail in Fig. 1b):

$$h_{x,1/2} = 2\sqrt{\ln 2}\sigma_x = 1.665\sigma_x \quad \text{and} \quad h_{y,1/2} = 2\sqrt{\ln 2}\sigma_y = 1.665\sigma_y \quad (6)$$

The average spot dimensions, σ_x and σ_y , are an estimation of the system performance (*see Note 6*).

4. Compute the number of proteins present in the mixture, m , from the 2-D EACVF value computed at the origin, i.e., $\Delta pI = 0$, $\Delta \log M_r = 0$, $C_{0,0}$, using the following equation (enlarged detail in Fig. 1b; *see Note 7*):

$$m = \frac{V_T^2(\sigma_h^2/a_h^2 + 1)\ln 2}{\pi h_{x,1/2}h_{y,1/2}C_{0,0}XY} = \frac{0.22 V_T^2(\sigma_h^2/a_h^2 + 1)}{h_{x,1/2}h_{y,1/2}C_{0,0}XY} \quad (7)$$

This computation requires the detection of the spots and the evaluation of their intensities (i.e., volumes) to compute the total volume, V_T , and the relative dispersion ratio of the spot maxima, σ_m^2/a_m^2 .

5. Detect the p_{\max} spot maxima ($h_{m,j}, \forall j = 1, p_{\max}$) present in the map using an algorithm based on the comparison of seven successive points for each dimension. In this approach, the maximum (fourth point) was detected when the first three values were increasing and the last three decreasing. Alternatively, the dedicated software, Melanie II [22], was employed to yield a more correct measure of the volumes $V_{m,j}$ of the detected p_{\max} spots.
6. In the computation of Eq. 7, the most critical parameter is the estimation of the protein abundance dispersion ratio (*see Note 8*).
7. Compute the degree of separation achieved in the map, the separation extent $\gamma = p_{\max}/m$, as the ratio between the total number of the detected spots, p_{\max} , and the number of proteins, m [11, 12, 17, 20, 23]. This value is usually lower than 1, as a consequence of spot overlapping present in real 2-D PAGE maps (*see Note 9*).

Table 1

Computation of the separation parameters on digitized 2-D maps generated by computer simulations (first–sixth rows) and reference 2-D maps retrieved from the SWISS-2D PAGE database (seventh–ninth rows)

m	σ_x	σ_y	σ_h^2/a_h^2	m_{est}	$e_{\text{rel}} (\%)$	$\sigma_{x,\text{est}}$	$\sigma_{y,\text{est}}$	σ_m^2/a_m^2	p_{max}	$\gamma (\%)$
200	0.026	0.0006	1	203 ± 14	7	0.026	0.0006	0.93	187	92
200	0.009	0.0002	1	196 ± 14	7	0.009	0.0002	0.97	184	94
500	0.026	0.0006	1	459 ± 21	5	0.026	0.0005	0.92	381	83
500	0.009	0.0002	1	485 ± 22	5	0.009	0.0002	0.97	451	93
750	0.009	0.0002	1	710 ± 26	4	0.009	0.0002	0.95	660	93
1000	0.009	0.0002	1	919 ± 3	3	0.009	0.0003	0.93	836	91
HEPG2 99	0.009	0.0002	1	100 ± 10	10	0.009	0.0002	0.99	99	99
DL-1 108	0.009	0.0002	1	104 ± 10	10	0.009	0.0002	0.99	103	99
PLASMA 626	0.009	0.0002	1	601 ± 24	4	0.009	0.0002	0.96	571	95

The true values of number of proteins (m) and spot shape (σ_x, σ_y) are compared to the estimated results ($m_{\text{est}}, \sigma_{x,\text{est}}$ and $\sigma_{y,\text{est}}$): $e_{\text{rel}}\%$ is the percentage relative error of m_{est} related to m . The theoretical value of protein abundance dispersion ratio, σ_h^2/a_h^2 , is compared with its experimental approximation, the maximum dispersion ratio, σ_m^2/a_m^2 . From the total number of detected spots, p_{max} , the degree of separation achieved γ was computed.

HEPG: hepatoblastoma-derived cell line (HEPG2_HUMAN) [19]

DL-1: colorectal adenocarcinoma cell line (DL-1) (DLD1_HUMAN) [19]

PLASMA and a human plasma (PLASMA_HUMAN) [19]

- As an example, the method was applied to synthetic 2-D maps describing the separation pattern usually present in 2-D PAGE gels and reference 2-D maps retrieved from the SWISS-2D-PAGE database (Table 1) [16]. The elliptical spot shape with $\sigma_x = 0.009$ pI and $\sigma_y = 0.0002$ logMr (second, fourth–ninth rows) corresponds to the common 2-D PAGE maps (*see Note 4*). Spot abundance (AM) was described by an exponential (E) distribution yielding $\sigma_h^2/a_h^2 = 1.0$ (fourth column).

The results obtained on the investigated maps (Table 1) show that the 2-D autocovariance function method gives a correct estimation of the mean spot shape (compare seventh and eighth columns with second and third columns) and the number of proteins present in the sample (compare fifth column with first column; *see Note 10*).

The obtained values of the degree of separation (γ , 11th column) show that if 500 proteins are present in a sample, only 83 % of them can be separated in the worst conditions corresponding to lower efficiency ($\sigma_x = 0.026$ pI and $\sigma_y = 0.0006$ logMr, third row) but separation increases up to 93 % with greater efficiency ($\sigma_x = 0.009$ pI and $\sigma_y = 0.0002$ logMr, fourth row).

3.4 Estimation of the Separation Pattern: Detection of Presence of Spot Ordered Sequences

1. Analyze the 2-D EACVF region describing long-term correlation (interdistance higher than $4\sigma_x$ and $4\sigma_y$, Fig. 1b) to extract information on the retention pattern of the 2-D maps. If the 2-D map exhibits some ordered structures formed by spots located at constant interdistances (Δx , Δy) repeated in different regions of the map, the 2-D EACVF plot shows well-shaped deterministic cones at the repeated Δx , Δy values (Fig. 3b shows the 2-D EACVF plot computed on the digitized map of Fig. 3a). The visual detection of such deterministic peaks permits to identify the existence of the sequences, singling them out from the signal complexity [16–18] (*see Note 11*). If the repeated interdistances (Δx , Δy) correspond to sequences of protein spots, the Δx and Δy values are related to the parameters b_x and b_y of the series (Eqs. 3a and 3b) and intensities of the 2-D EACVF cones to the number of terms of spot sequence, n_{\max} (Eq. 4) [16].
2. A common feature of the 2-D PAGE maps of biological samples is the presence of trains of protein spots that are consistent with the separation of protein isoforms differing in a constant change in amino acid charges or molecular weight or produced by co- and post- translational modifications (PTMs) such as glycosylation and phosphorylation [1, 24] (*see Note 12*).
3. As an example, the 2-D PAGE map of a hepatoblastoma-derived cell line (HEPG2_HUMAN from SWISS-2-D PAGE database 19) was investigated (Table 1, seventh row and Fig. 3a). The 2-D EACVF plot computed on the map (Fig. 2b) shows some well-defined deterministic cones that identify some interdistance repetitiveness present in the map (*see Note 13*).

This behavior is more clearly shown by projecting the 2-D EACVF values along each separation axis, i.e., ΔpI and ΔMr (black line in insets in Fig. 3b). For comparison, the figures show the plots of the 2-D EACVF computed on a simulated map containing the same number of components ($m = 99$) with a disordered retention pattern (red line). Well-defined deterministic peaks are evident along the pI axis (inset at the top): one peak at $\Delta pH = 0.2$ that is consistent with protein isoforms generated by acetylation and another at $\Delta pH = 0.5$, which is consistent with protein alkylation [1, 2, 24]. Also along the second dimension separation axis (inset at the bottom), some deterministic peaks are evident due to repeated increments of the protein molecular masses, i.e., at ΔMr value of 160 and 330 Da, which can be related to protein acylation or glycosylation (*see Note 14*).

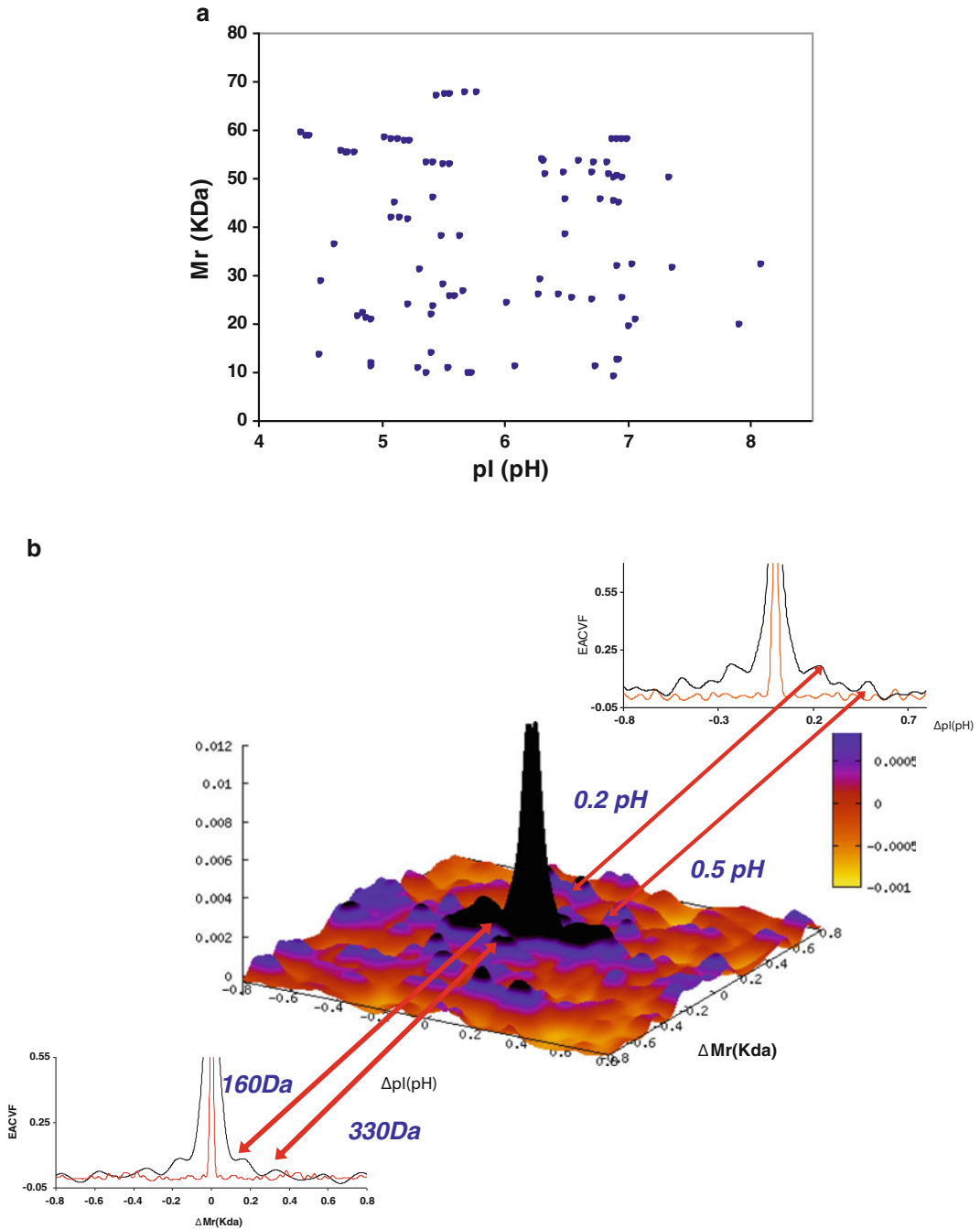


Fig. 3 2-D EACVF method on a real 2-D PAGE map for identification of train of spots. **(a)** Digitized 2-D PAGE map of a hepatoblastoma-derived cell line (HPG2) from SWISS-2D PAGE database. **(b)** Plot of the 2-D autocovariance function computed on the map. Enlarged insets: 2-D EACVF values over the ΔpI (*top*) and ΔMr (*bottom*) separation axes - comparison between the 2-D EACVF computed on the HEPG2 map (*black line*) and on a simulated map containing the same number of components with a disordered retention pattern (*red line*)

4 Notes

1. The theoretical background of this approach is that a multi-component 2-D map is considered as a series of 2-D spots with a random distribution of position and height. For sake of simplicity, here we assume that the spots are modeled by bidimensional Gaussian peaks; thus, the signal is expressed as:

$$f(x, y) = \sum_{i=-\infty}^{\infty} h_i \exp \left[-\frac{(x - x_{0,i})^2}{2\sigma_x^2} - \frac{(y - y_{0,i})^2}{2\sigma_y^2} \right] \quad (8)$$

where h_i is the random height, $x_{0,i}$ and $y_{0,i}$ the random positions of the spots, and σ_x and σ_y the widths of the elliptic 2-D spots along the separation axes x and y .

2. There are theoretical and experimental evidences that in 2-D PAGE maps of biological samples, the disordered pattern is the most probable distribution described by an exponential function, as a consequence of the mixture complexity itself. It derives from the combination of two independent distributions along the separation axes and yields a fully random pattern with spot overcrowding in the central region of the map [1, 2, 11–14, 23, 24].
3. Usually, the real 2-D maps exhibit a combination of general disordered spots with superimposition of some ordered structures. They are formed by spots located at constant interdistances (ΔpI , $\Delta \log Mr$) repeated in different regions of the map, or ordered sequences of spots, such as the case of trains of spots, which correspond to protein isoforms showing a monodimensional shift of the spots to more basic/acid pI values.
4. As an example, two specific cases were considered. The elliptical spot shape with $\sigma_x = 0.009$ pI and $\sigma_y = 0.0002$ $\log Mr$ (second, fourth–ninth rows in Table 1) describes a common 2-D PAGE map where a tissue homogenate sample (*ca.* 1 mg of total protein) is loaded on a standard gel size of 18-cm strip of broad pH range (pH 3–7) and analyzed with standard (1 mm) scanner resolution. Spots with $\sigma_x = 0.026$ pI and $\sigma_y = 0.0006$ $\log Mr$ correspond to the same 2-D PAGE map analyzed with lower (3 mm) scanner resolution (first and third rows in Table 1).
5. A crucial point of the numerical calculation of 2-D EACVF is computing each node using the same number of points to assure the same degree of precision. For this purpose, use the cyclic calculation procedure by merging the beginning and the end of the separation axes by using negative k or l indices [15]. This means that the 2-D map is handled as it is

wrapped around itself and the right side of the separation space is continuously linked to the left one [25].

6. The estimation of the system performance is the basis for selecting proper experimental conditions in order to optimize the efficiency of the 2-D separation, i.e., improving the gel structure, increasing the gel size, or exploiting narrower and narrower pH gradients in the first dimension IEF separation or both. Another way for reducing ΔpI and $\Delta \log Mr$ is increasing scanner resolution, i.e., from standard (1 mm) to high (<0.5 mm) [16]. Moreover, high σ_x and σ_y may be diagnostic for overloading effects revealing that an excess of sample has been loaded on the gel.
7. It must be noted that the obtained data are a statistical estimation of the corresponding parameters, that means that they are the average statistical attributes of the separation, without any specific information on the spot of each protein.
8. The protein abundance dispersion ratio σ_h^2/a_h^2 must be approximated to the spot maximum dispersion ratio (σ_m^2/a_m^2) computed from the spot maxima detected in the map. In fact, spot maxima are the only experimentally accessible values that exactly correspond to protein abundance only if all the proteins are carefully separated in the absence of spot overlapping. σ_m^2/a_m^2 is computed from the peak maximum values using the average peak maximum abundance, where σ_m^2 and its standard deviation σ_m^2 . Alternatively, it can be calculated as the relative dispersion ratio of spot volume, σ_v^2/a_v^2 , using the volumes $V_{m,j}$ of the detected p_{max} spots measured with the dedicated software, Melanie II.

The results obtained from computations on the simulated maps, where σ_h^2/a_h^2 value is a priori defined, clearly show that the maximum dispersion ratio, σ_m^2/a_m^2 , computed by the developed algorithm is a correct estimation of the theoretical value σ_h^2/a_h^2 (compare values in ninth and fourth columns in Table 1).

9. It must be underlined that m is usually an experimentally inaccessible parameter of the map, since the total number of detected spots, p_{max} , is usually lower than m . This is a consequence of uncompleted separation process generating protein overlapping [12, 20, 23].
10. The parameter m can be estimated with a precision \sqrt{m} , since it is computed by using the Poissonian statistics. For all the studied maps, an accurate estimation with an error lower than 10 % was obtained (expressed as percentage relative error, $e_{rel}\%$, sixth column in Table 1), even if the number of proteins considered is low and the theoretical model describing the distribution of the positions is far from being thoroughly known [16].

11. The ACVF approach has a specific ability in recognizing structured distribution in the complex separation space, since the autocovariance function is able to cancel the effect of the randomness of spot positions, while it amplifies the recursivity of the repeated interdistances. In fact, it is based on the Fourier analysis, which has been long used to identify periodicity and order, deeply buried in signals [18, 25].
12. Most protein isoforms clearly show different pI values without significant mass variations. A case in point is deamidation, that yields in 2-D map trains of spots horizontally oriented in respect to the mass axis (indeed the 1-Da difference typical of $\text{Asn} \rightarrow \text{Asp}$ and $\text{Gln} \rightarrow \text{Glu}$ transitions was found) and positioned at lower pI values in respect to a potential parental spot. Moreover, such extensive trains of spots seemed to be particularly frequent in sera and less than this in other human tissues (e.g., liver) [1, 2, 26].

The identification of protein post-translational modifications is quite an important aspect of proteomics, since it has been well established that PTMs occur on many proteins and are of extreme biological importance. In fact a vast number of proteins have been shown to be post-translationally regulated by a variety of different PTMs, i.e., change in enzymatic activity, ability to interact with other proteins, subcellular localization, targeted degradation, etc. [1, 24, 26].

13. The plot of Fig. 3a clearly shows that 2-D EACVF exhibits a C_2 symmetry: correlations in positions $(\Delta x, \Delta y)$, $(-\Delta x, \Delta y)$, and $(\Delta x, -\Delta y)$ are equal to those in $(-\Delta x, -\Delta y)$, that means that both positive and negative ΔpI shifts give the same 2-D EACVF values.
14. It must be noted the 2-D EACVF method provides high sensitivity in detecting order: for example, it has been reported that it is able to detect the presence of only sevenfold repetitiveness hidden in a random pattern of 200 proteins [17].

References

1. Rotilio D, Della Corte A, D'Imperio M, Coletta W, Marcone S, Silvestri C, Giordano L, Di Michele M, Donati MB (2012) Proteomics: bases for protein complexity understanding. *Thromb Res* 129:257–262
2. Simula MP, Notarpietro A, Toffoli G, De Re V (2012) 2-D gel electrophoresis: constructing 2D-gel proteome reference maps. *Methods Mol Biol* 815:163–173, *Functional Genomics: methods and Protocols*, 2nd Edition, Book Series
3. Clement CC, Aphkhasava D, Nieves E, Callaway M, Olszewski W, Rotzschke O, Santambrogio L (2013) Protein expression profiles of human lymph and plasma mapped by 2D-DIGE and 1D SDS-PAGE coupled with nanoLC-ESI-MS/MS bottom-up proteomics. *J Proteomics* 78:172–187
4. Szabo Z, Szomor JS, Foeldi I, Janaky T (2012) Mass spectrometry-based label free quantification of gel separated proteins. *J Proteomics* 75:5544–5553
5. Colignon B, Raes M, Dieu M, Delaive E, Mauro S (2013) Evaluation of three-dimensional gel electrophoresis to improve

- quantitative profiling of complex proteomes. *Proteomics* 13:2077–2082
6. Nakano K, Tamura S, Otuka K, Niizeki N et al. (2013) Development of a highly sensitive three-dimensional gel electrophoresis method for characterization of monoclonal protein heterogeneity. *Anal Biochem* 438:117–123
 7. Lee BS, Gupta S, Morozova I (2003) High-resolution separation of proteins by a three-dimensional sodium dodecyl sulfate polyacrylamide cube gel electrophoresis. *Anal Biochem* 317:271–275
 8. Moche M, Albrecht D, Maass S, Hecker M, Westermeyer R, Buttner K (2013) The new horizon in 2D electrophoresis: new technology to increase resolution and sensitivity. *Electrophoresis* 34:1510–1518
 9. Li F, Seillier-Moisewitsch F, Korostysheyskiy VR (2011) Region-based statistical analysis of 2D PAGE images. *Comput Stat Data Anal* 55:3059–3072
 10. Marengo E, Robotti E (2012) A new algorithm for the simulation of SDS 2D-PAGE datasets. *Methods Mol Biol* 869:407–425
 11. Pietrogrande MC, Marchetti N, Dondi F, Righetti PG (2003) Spot overlapping in 2D-PAGE maps. Relevance to proteomics. *Electrophoresis* 24:217–221
 12. Campostrini N, Areces L, Rappsilber J, Pietrogrande MC, Dondi F, Pastorino F, Ponzoni M, Righetti PG (2005) Spot overlapping in two dimensional maps: a serious problem ignored for much too long time! *Proteomics* 5:2385–2395
 13. Rabilloud T, Vaezzadeh AR, Potier N, Lelong C et al. (2009) Power and limitations of electrophoretic separations in proteomics strategies. *Mass Spectrom Rev* 28:816–843
 14. Rabilloud T (2013) When 2D is not enough, go for an extra dimension. *Proteomics* 13:2065–2068
 15. Marchetti N, Felinger A, Pasti L, Pietrogrande MC, Dondi F (2004) Decoding two-dimensional complex multicomponent separations by Autocovariance Function. *Anal Chem* 76:3055–3068
 16. Pietrogrande MC, Marchetti N, Tosi A, Dondi F, Righetti PG (2005) Decoding 2-D PAGE complex maps by Autocovariance function: a simplified approach useful for proteomics. *Electrophoresis* 26:2739–2748
 17. Pietrogrande MC, Marchetti N, Dondi F, Righetti PG (2006) Decoding 2-D PAGE complex maps by Autocovariance function: relevance to proteomics. *J Chromatogr B* 833:51–62
 18. Dondi F, Pietrogrande MC, Marchetti N, Felinger A (2008) Decoding complex 2-D separations in multidimensional liquid chromatography. In: Cohen SA, Schure MR (eds) *Theory and applications in industrial chemistry and the life science*. John Wiley & Sons, New York, pp 59–90, ISBN: 978-0-471-73847-3
 19. Human 2-D PAGE database of *Danish Centre for Human Genome Research*, <http://www.biobase.dk/cgi-bin/celis>
 20. Pietrogrande MC, Marchetti N, Dondi F, Righetti PG (2002) Spot overlapping in 2D-PAGE separations: a statistical study of complex protein maps. *Electrophoresis* 23:283–289
 21. Press WH, Teukosky SA, Vetterling WT, Flannery BP (1986) *Numerical recipes in Fortran*. Cambridge University Press, Cambridge, UK. ISBN 978-0521309585
 22. Melanie II, Geneva Bioinformatics, GeneBio S. A., <http://www.genebio.com>.
 23. Davis JM, Giddings JC (1984) Origin and characterization of departures from the statistical model of component-peak overlap in chromatography. *J Chromatogr* 289:277–298
 24. Righetti PG, Candiano G (2011) Recent advances in electrophoretic techniques for the characterization of protein biomolecules: A poker of aces. *J Chromatogr A* 1218:8727–8737
 25. Felinger A (1998) *Data analysis and signal processing in chromatography*. Elsevier, Amsterdam. ISBN 978-0444820662
 26. Corfe BM, Evans CA (2014) Are proteins a redundant ontology? Epistemological limitations in the analysis of multistate species. *Mol Biosyst* 10:1228–1235

2-D PAGE Map Analysis

Methods and Protocols

Marengo, E.; Robotti, E. (Eds.)

2016, XVI, 331 p., Hardcover

ISBN: 978-1-4939-3254-2

A product of Humana Press