Automatic analysis of two-dimensional electrophoresis gel images for applications in proteomics

I. De Mitri, [1] [2] G. De Nunzio, [3] [2] M. Maffia, [4] S. Maglio, [3] [2] and G. Mercurio [5]

[1]Dipartimento di Fisica, Università del Salento, Lecce, Italy

[2]Istituto Nazionale di Fisica Nucleare, Sezione di Lecce, Italy

[3]Dipartimento di Scienza dei Materiali, Università del Salento, Lecce, Italy

[4]Dipartimento di Scienze e Tecnologie Biologiche e Ambientali, Università del Salento, Lecce, Italy

[5]Unità di Sanità Elettronica, CNR Istituto Tecnologie Biomediche, Roma, Italy

Considerable progress has been made in the human biology, since the discovery of DNA structure: the "human genome" provided much information on the sequences of individual genes, and new molecular biology techniques (PCR, DNA microarray) allowed the analysis of gene expression at the "transcriptome" level (messenger RNA pool in a cell). However, in recent years, the scientific community interest has shifted towards the study of the structure and function of proteins. This has happened for various reasons, including the fact that the static nature of the genome cannot describe the dynamics of cellular processes [1]. Proteomics is the science that studies the proteome, the proteic expression of the genome [2,3]. Cell proteome is extremely complex, and is composed of several thousand proteins. Two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) is widely used as a standard method to separate and display proteins in a tissue or compound up to a theoretical limit of 10^4 proteins simultaneously. This technique combines the resolution power of isoelectrofocalization (IEF), which distinguishes proteins by their isoelectric point, with SDS-PAGE (sodium-dodecyl-sulphate PAGE), in which proteins are separated according to their weight and molecular size.

Our group is currently developing software algorithms for the automatic analysis of images obtained by 2D-PAGE gel optical scanning (see Fig.1). The goal is the reduction of human intervention in the analysis process (which consists in protein recognition by image comparison) and the possibility to make automatic quantitative assessments on the analyzed image. The human-guided process is currently quite slow, operator-dependent, and error-prone. This note is a brief summary of our work, more details and preliminary results being reported in [4–8].

Several (semi)automatic analysis techniques for 2D-PAGE images are available in the literature (e.g. [9]), but the problem complexity still demands for more complete solutions. Our procedure currently currently consists in several steps, coded in Matlab and C/C++. As a first step, image noise is reduced, in order to limit false positive detection and protein misidentification. Proteins appear as dark spots on a quite light background (see Fig.1), so the next step is the search for local minima. This step requires the choice of a threshold value, which would exclude irrelevant minima. We have bene as conservative as possible, so that (almost) no significant spots are neglected. The drawback is the presence of a large number of false positives, which must be later recognized and eliminated. Unfortunately, not all of the protein spots are identified by the minima search procedure. Sometimes two or more of them actually merge and are counted as just one spot, because one is deeper and the others are only 'shoulders' of the main one. The watershed transform is thus applied, which partitions the gel into basins: each basin contains a single (recognized) minimum, but (as just pointed out) can also include more spots close to the main one.

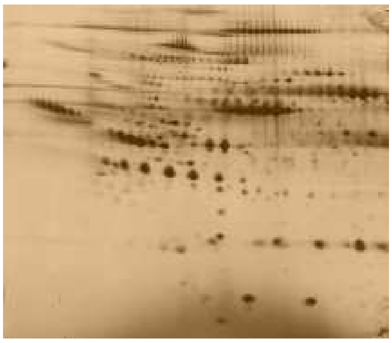

Figure 1. A two-dimensional electrophoresis gel image. In principle, each spot corresponds to a given protein in the analyzed sample (see text).

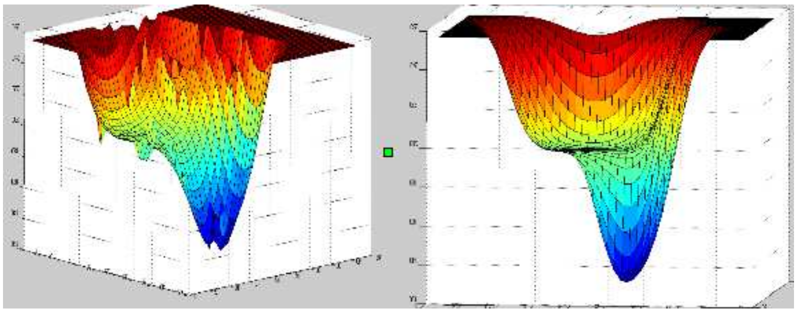

Figure 2. On the left: the grey level profile of a region with two spots in a basin (3D view). On the right: the same region as parameterized with our model.

We therfore use some images already analyzed by biologists as "atlas". We search the most important spots both in the atlas and in the analyzed image. A suitably developed algorithm is then applied in order to search for the most important spots (by choosing the deeper and larger ones), obvioulsy neglecting multiple spots joined together and sidelong ones. The coordinates of the most important atlas spots are registered with the analyzed image by using the Robust Point Matching algorithm.

We then project all the atlas spots to the image by the transformation deduced in the previous step. Each basin is then used as a region of interest (ROI) in which the shape of the spot (or spots) is fitted through a χ^2 minimization procedure. We use a model, derived from [10], with six parameters for each spot and one parameter for the background value. This model considers that the spot can be asymmetrical, that there can be comigrating spots in complex regions, and that spots can be saturated.

Equation (1) represents the parameterization model of a single spot in a basin.

$$\begin{aligned} &C\left(x,y\right)=B+\\ &+\frac{1}{2}C_0\left[erf\left(\frac{a'+r'}{2}\right)+erf\left(\frac{a'-r'}{2}\right)\right]+\\ &+\frac{C_0}{r'}\sqrt{\frac{1}{\pi}}\left\{\exp\left[-\frac{\left(a'+r'\right)^2}{4}\right]-\exp\left[-\frac{\left(a'-r'\right)^2}{4}\right]\right\} \end{aligned} \tag{1}$$

Here $C(x,y)$ is the concentration of the diffusing substance at the point identified by the coordinates x and y, B is the background level,C_0 is the initial concentration of the diffusing substance, a' is a parameter bound to the spot size. The meaning of r' is in (2):

$$r^{'}=\sqrt{\frac{\left(x-x_0\right)^2}{D_x^{'}}+\frac{\left(y-y_0\right)^2}{D_y^{'}}} \tag{2}$$

where $D_x^{'}$ and $D_y^{'}$ are diffusion coefficients, x_0 and y_0 are the coordinates of the position of the spot minimum. In case of two or more spots in the basin, we use a generalization of (1). The coordinates of the transformed spots are used as initialization parameters in our fit. In Fig.2 we can see the application of our model to a basin containing two spots.

According to our tests, some commercial software tools (such as Melanie or PD-Quest) recognize, in the analyzed images, roughly 2000 spots, with 90% sensitivity and 1200 false positives (later rejected by biologists by hand in several hours). Tests on our images are in progress. Anyway, before coregistration, we recognize in average 1100 spots, with 80% sensitivity and 400 false positives. After coregistration, the number of false positives is significantly reduced. Now we are working on the parameterization/fitting step to improve our encouraging results.

REFERENCES

1. G. Chambers, L. Lawrie, P. Cash et al., Journal of Pathology 2000; 192 (3): 280-288.
2. V. C. Wasinger, S. J. Cordwell, A. Cerpa-Poljak et al., Electrophoresis 1995; 16 (7): 1090-1094.
3. M.R. Wilkins, J. C. Sanchez, A. A. Gooley et al., Biotechnology & Genetic Engineering Reviews 1996; 13: 19-50.
4. G. De Nunzio et al., 21st IEEE International Symposium on Computer-Based Medical Systems, Jyvaskyla, Finland, 2008.
5. G. De Nunzio et al., BIOTECHNO 2008: International Conference on Biocomputation, Bioinformatics, and Biomedical Technologies, Bucharest, Romania, 2008.
6. S. Maglio et al., BMC Bioinformatics Special Issue on NETTAB 2008 workshop, Como, Italy, 2008.
7. S. Maglio et al., XCIV Congresso Nazionale SIF - Genova, Italy, 2008.
8. G. De Nunzio et al., MIC 2008: IEEE International Conference, Dresden, Germany, 2008.
9. W. Van Belle et al., BMC Bioinformatics 2006, 7:198.
10. E. Bettens et al., Electrophoresis 1997 Vol. 18 pp. 792-798