4

Statistical Methods for Proteomics

Françoise Seillier-Moiseiwitsch, Donald C. Trost, and Julian Moiseiwitsch

1. Introduction

What is Proteomics? The term *proteome* denotes the PROTEin complement expressed by a genOME or tissue. While the genome is an invariant feature of an organism, the proteome depends on its developmental stage, the tissue considered, and environmental/experimental conditions. There are more proteins in a proteome than genes in genome (which is particularly true for eukaryotes). For instance, there are several ways to splice a gene to generate messenger ribonucleic acid (mRNA). Furthermore, proteins can undergo posttranslational alterations such as truncation at the amino- (N)- and carboxy (C)-terminus and addition of saccharide or phosphate groups.

Two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) is currently the only method able to separate thousands of proteins. Mammalian cell samples, for example, exhibit more than 2000 proteins **(Fig. 1)**. Two coordinates characterize each protein: its isoelectric point and its molecular mass.

For the first dimension, proteins are focused electrophoretically along a pH gradient. Their movement stops when they reach a position at which they have no net charge (i.e., their isoelectric point). For the second dimension, proteins are soaked in sodium dodecyl sulfate (SDS) so that all proteins acquire the same charge density. They are then separated orthogonally by electrophoresis on a polyacrylamide gel according to their molecular weight. Under carefully controlled experimental conditions, these two dimensions, the isoelectric point and molecular mass, are independent. The separated proteins are then stained with fluorescent dyes so that they are readily detectable. The image of the displayed proteins defines the proteome. This image is digitally scanned into a database for storage ***(1–4)***.

From: *Methods in Molecular Biology, vol. 184: Biostatistical Methods*
Edited by: S. W. Looney © Humana Press Inc., Totowa, NJ

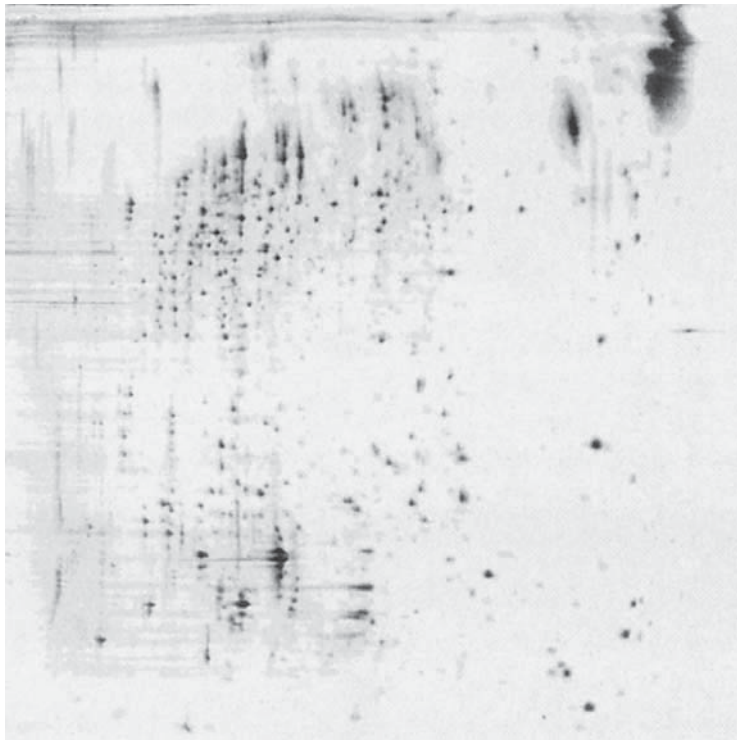

Fig. 1. 2D-PAGE image of kidney tissue sample.

This technology still presents many challenges to the experimenter. Gel quality must remain constant from day to day. Gel reproducibility from laboratory to laboratory cannot always be guaranteed. Membrane proteins necessitate a special protocol, otherwise they are underrepresented: losses are due to hydrophobic interactions between these proteins and the gel. The experimental procedure must ensure the removal of nucleic acids as these can cause streaks and artifactual migration of some proteins. Finally, the abundance threshold for detection on a gel is still to be determined.

2D-PAGE allows the systematic analysis of proteins for any disease and in any biological system. As proteins are responsible for phenotypes, they are the direct targets for therapeutic agents. Therefore, this technology has great potential for aiding in the drug-discovery process ***(5)*** and in medical diagnosis. The specific areas of applications are:

1. Treatment monitoring. 2D-PAGE has been used to assess treatment effects on tumors and to study overall protein expression following hormone therapy.
2. Identification of disease-specific proteins. In cancer studies, 2D-PAGE is utilized to compare protein expression in normal and cancerous tissues, therefore identifying candidate targets for drugs. It thus serves a purpose similar to that of

genomic microarray analysis where the goal is to identify clusters of disease-specific mRNAs. The disadvantage of the latter approach is that levels of a specific mRNA and the corresponding protein may not be correlated.
3. Target validation and signal transduction studies.
4. Drug mode-of-action studies.
5. Drug toxicology studies.

2. Technology Background

2.1. Electrophoresis

Electrophoresis is the process of separating a mixture of electrically charged molecules by applying an electric field. This charge is either due to charged groups on the molecules themselves (*see* **Subheading 2.1.1.**) or associated with a coating of charged molecules (*see* **Subheading 2.1.2.**). Electrophoresis is commonly used in clinical chemistry to separate macromolecules, such as proteins, in tissue samples and bodily fluids for identification or quantification. A brief overview of the physics of electrophoresis ***(6)*** is presented here to link the technology with the analytical methods.

Molecules can be separated via electrophoresis according to either their innate charge or their molecular weight, or both as in two-dimensional gels. Migration is also influenced by the shape of the molecule, the strength of the electric field, the ionic properties and pH of the electrophoresis buffer, and the temperature of the system. The electric field acts on the molecule according to Coulomb's law: the force on the object is proportional to its net charge (z) and the strength of the field (E). The velocity (v) of the molecule is equal to the force divided by the strength of the electric field. The potential for movement within the gel, or electrophoretic mobility (M), is then defined as v/E. With the assumption that the molecule is spherical,

$$M = \frac{ze}{6\pi\eta r}$$

where e is the electrostatic constant, η the medium viscosity, and r the macromolecular radius.

Several types of electrophoresis have been utilized since Tiselius ***(7)*** developed the first method, moving-boundary electrophoresis, to study proteins. Moving-boundary electrophoresis has largely been replaced by zonal electrophoresis ***(8)***. This method uses a thin band, or zone, of macromolecule solution placed in a semisolid matrix such as a gel. After exposure to the electric field, the sample separates into bands of molecules with similar electrophoretic mobility. Usually gels are composed of agarose or polyacrylamide in varying percentages. However, starch and cellulose acetate have also been used and continue to have their applications. Some gels also act as a molecular sieve that separates molecules based on size.

When the electrophoresis medium comes into contact with water, hydroxyl ions adsorb to the surface, creating a stationary layer of negative charge. This layer attracts a cloud of positive ions. When the current is applied, the flow of positive ions can slow protein movement or reverse its direction. This unwanted change in mobility is called *endosmosis*. This effect tends to be strong in cellulose paper, cellulose acetate, and agarose gel, but is minimal in starch gel and polyacrylamide gel. For this reason, polyacrylamide provides better resolution, and is used when proteins of similar molecular weight need to be differentiated.

2.1.1. Isoelectric Focusing

Separating molecules by their inherent charge using electrophoresis is termed *isoelectric focusing*. In isoelectric focusing, the electrophoresis medium is constructed so that a pH gradient exists between the electrodes. Proteins have many ionization sites that are pH dependent. Such a molecule, that can be either positively or negatively charged, is called a *zwitterion*. The gel buffer must have the ability to both donate and accept hydrogen ions, which allows the charge of a zwitterion to change as it moves through the pH gradient (such a buffer is called an *ampholyte*). For each protein, there is a pH at which the number of positively charged sites is equal to the number of negatively charged sites, giving a net charge of zero. At this point in the pH gradient, the electrophoretic mobility is zero as the protein has no net pull to move. This is called the *isoelectric point* (pI).

2.1.2. SDS-PAGE

A protein can be denatured by heating it in the presence of SDS, an ionized detergent. The heat causes the protein to unfold into a long strand, and allows the detergent to form a *micelle* (molecular cage) around it. This micelle behaves like a long rod with a surface charge proportional to its length, which is proportional to the protein's molecular weight and much larger than any net charge inherent in the macromolecule itself. As the charge is now proportional to the molecular weight, the amino-acid sequence loses its importance in determining migration, and mobility is then inversely proportional to the radius of the molecule. This is the basis for SDS-gel electrophoresis. By equating the volume of a sphere of radius r with the volume of a rod of length L and radius b, the spherical frictional coefficient of the rod can be derived ***(6)***, and the electrophoretic mobility of a rod-shaped molecule is

$$\frac{c\left[\log\left(\frac{L}{b}\right) - 0.30\right]}{3\pi\eta}$$

where c is a proportionality constant relating micellar charge to molecular length. This implies that very long molecules have similar mobilities because of the flatness of the logarithm function.

The electrophoresis buffer carries the current and determines the pH of the medium. A buffer is an ionic solution of a weak acid (or a weak base) plus its salt, and functions to maintain a constant pH. The relationship is described by the Henderson–Hasselbalch equation:

$$\mathrm{pH} = pK_a + \log_{10}\left(\frac{\mathrm{salt}}{\mathrm{acid}}\right)$$

where pK_a is a constant that depends on the acid (or base). When pH is near pK_a, the curve is relatively flat over a considerable range of the ratio. A constant pH is important because the conformation of the macromolecules can be drastically affected by a change in pH: proteins return to their folded configuration. As a result, the molecule is no longer a rod and moves more slowly through the gel than would be expected based on its molecular weight alone. As various proteins return to their natural configuration at different pH values, a stable buffer is important for reliable SDS-PAGE.

The ionic strength of the buffer impacts mobility. At high ionic concentration, the ions hinder the movement of the protein by forming a cloud around the macromolecule. Different buffers (e.g., with varying concentration of Tris) are appropriate depending on the size of the molecule. If smaller molecules are to be separated, higher buffer concentrations are used to improve resolution, while with larger molecules the buffer concentration is lowered to reduce the time required to run the gel.

2.1.3. 2D-PAGE

Kenrick and Margolis ***(9)*** did the initial work on two-dimensional electrophoresis by combining isoelectric focusing with SDS-gel electrophoresis. Subsequently, Klose ***(10)***, O'Farrel ***(11)***, and Scheele ***(12)*** each published applications of this method which form the basis for current techniques of 2D-PAGE. Isoelectric focusing is applied in one direction to separate proteins based on their pI. The gel is then soaked in SDS, and the resultant bands are exposed to an electric field, perpendicular to the first one. With this technique, several thousand proteins can be isolated on a single gel. Considerable effort has been invested to improve the resolution and reproducibility of these gels to maximize the number of proteins detected and to reduce the within- and between-laboratory variability.

2.2. Gel Preparation

In polyacrylamide gels, the acrylamide monomer and a crosslinking agent such as bisacrylamide are mixed and react in the presence of other reagents to

form an acrylamide polymer ***(13)***. The rate at which molecules pass through a gel is determined by the pore size of the gel. For a given pore size, larger molecules will travel more slowly. The pore size of the gel is determined by the percentage of crosslinker it contains. Gels are characterized by the total percentage of acrylamides (linear and crosslinker) and the percentage of crosslinker. In addition to a separating gel, a stacking gel is used to concentrate the proteins into a stack of very thin bands before they reach the separating gel. The proteins are loaded onto a gel by placing them in a well that is several millimeters deep. If they are immediately placed onto the separating gel, proteins at the top of the well have significantly further to travel before they enter the gel than those at the bottom of the well. To account for this, the stacking gel concentrates all proteins into a very thin band (<1 mm thick). A spacer gel is sometimes used between the stacking and the separating gels when large concentrations of proteins are to be separated.

An early method for isoelectric focusing used carrier ampholytes ***(8,14)***. These small molecules, polyaminocarboxylic acids, rapidly migrate to a location in the gel that corresponds to the isoelectric point of the ampholyte. The mixture of ampholytes with varying pI sets up a pH gradient before the proteins have migrated very far. The proteins then move to their respective isoelectric point. This method has several drawbacks. After stabilization, the ampholytes tend to drift toward the cathode, causing the gradient to vary. These gradients are also distorted by high salt or high protein concentrations. Furthermore, the gels become stretched when extruded from the glass tubes used to hold the gel. An alternative method, the immobilized pH gradient (IPG) method, was developed by Bjellqvist et al. ***(15)***. For this method, the pH gradient is created by covalently binding molecules to the acrylamide gel, allowing this gradient to be tailored to the problem. The gradient is typically a linear, step, or sigmoidal function of location. These gels are now commercially available, thereby reducing much of the variability in the gradients.

SDS-PAGE can be run either vertically or horizontally. Neither method offers an advantage with respect to reproducibility and both require stacking gels. By convention, PAGE is usually performed vertically and agarose gel electrophoresis horizontally. With precast gels, this is purely for historical reasons. However, acrylamide does not polymerize completely in contact with air. Consequently, when these gels are poured in the laboratory (as opposed to utilizing precast commercial gels), air needs to be excluded from the surface, usually by placing a layer of water or mineral oil over the unset acrylamide. As it is convenient to have as small an amount of this liquid as possible, the gel is poured vertically and a layer of water is gently instilled on top. By contrast, agarose gels set by cooling at room temperature; the molten gel is poured into a horizontal mold so the entire surface can cool rapidly by heat radiating into

the air. As commercial gels are usually poured in an inert atmosphere and are ordered set, it is unimportant whether they are run vertically or horizontally.

2.3. Sample Preparation

Protein samples can be obtained from any cell, tissue, or protein-containing fluid. Special preparations may be required to break open cells (*lysis*) or to extract proteins from membrane structures. Most preparations include a buffer (to control pH and provide electrolytes for the current), detergents and chaotropic agents (to denature the protein and separate monomeric subunits), reducing agents (to break disulfide bonds), and a tracking dye (so that the progress of the electrophoresis can be monitored). If the sample is contaminated with nucleic acids, enzymes called *endonucleases* may be necessary to digest them, as nucleic-acid contamination can alter the electrophoretic characteristics of certain proteins.

2.4. 2D Gel Electrophoresis Procedure

In the first dimension, the proteins are separated by charge. For example, a gel strip, a few millimeters thick and containing a nonlinear (sigmoidal) immobilized pH gradient (3.5–10.0), can be made in the laboratory or purchased commercially for this purpose ***(16)***. The gel is connected to an electric current via the appropriate electrophoresis apparatus. After the sample is placed at the cathode end, the electric field is applied. The voltage is linearly increased from 300 to 3500 V over 3 h, then left at 3500 V for 3 h, and then at 5000 V for up to 100 kVh. After the run, the disulfide bonds are reduced chemically to prevent them holding the protein in a folded structure.

A vertical slab gel is typically used for the second dimension ***(17)***. The IPG strip is trimmed at the ends and placed in a solution at the top of the gel where they fuse. The gel is run at 40 mA with 100–400 V at 8–12°C for 5 h. A horizontal apparatus may be less efficient in transferring the protein from the first gel to the second gel, especially if the fusion is incomplete. Low-concentration proteins can be lost in the transfer.

2.5. Gel Staining and Scanning

Radiolabeling is the most sensitive method for localizing proteins on gels ***(14)***. This involves the binding of radioactive isotopes to the macromolecules and detecting, usually with film or phosphorescent screen, the radiation produced. With this method proteins can be detected at 20 parts per million. Chemoilluminescence and chemoilluminescence silver staining are alternative nonradioactive methods that are 100 times more sensitive than an organic dye such as Coomassie brilliant blue. Other stains include amido black, Ponceau S, and bromophenol blue. The silver staining method is started immediately after

the second run. The gel is washed, bathed in a soaking solution for several hours, and chemically treated. Then the staining solution is applied for 30 min, washed, and developed. Fluorescent stains can also be used and require a much shorter processing time because the fluorescence peaks after a relatively short time.

A laser densitometer is used to detect the concentration of staining. Each stain has a specific wavelength at which it emits the maximum signal. Improperly adjusted densitometers may produce peaks that are off the scale or eliminate short peaks when the background is improperly subtracted. This creates right and left censoring, respectively. The integral of each peak gives a quantity proportional to the abundance of the protein. Unfortunately, because the relationship between the intensity of staining and the protein concentration depends on the interaction between the protein and staining agent, in general, absolute concentrations cannot be calculated.

2.6. Spot Identification

Proteins can be identified by a number of methods ***(8,14)***. These include the determination of amino-acid composition and peptide mass fingerprinting. With several methods the individual proteins need to be transferred from the gel to another medium. Electroblotting to polyvinyldifluoride (PVDF) membranes is frequently used. Blotting can also be performed by semidry electrophoresis, vacuum, or capillary action. The advantage of semidry blotting is that it takes less than half an hour compared to other transfers that take several hours. The PVDF membranes are stained to locate the proteins, which are removed with a razor blade for identification. A minimum of 250 ng of protein per spot is required for identification. When analysis is performed using high-performance liquid chromatography (HPLC), the distribution of the amino acids, the pI, and the molecular weight are used to identify the protein from databases using a least-squares distance metric. An automated sequencer can only identify one protein per day while the HPLC method is 5–10 times faster. For peptide mass fingerprinting, the proteins are digested in the PVDF membrane using enzymes and analyzed using mass spectroscopy. The mass spectrum is then matched with a library of peptides. The ultimate goal for protein identification is to use standard maps overlaid on 2D-PAGE images, but these images require resource-intensive methods and sophisticated computer algorithms to develop the maps.

2.7. Other Sources of Variation

Protein separation and identification require many technical steps such as those described previously. Every step leads to (possibly high) variation in the output if not properly performed. A few additional sources of variation are described here:

1. Buffers are good growth media for microorganisms, which contain abundant protein. These buffers need to be stored in tight containers and refrigerated to inhibit growth. Periodically, they need to be replaced.
2. Care must be taken in sample collection, handling, and storage to avoid sample contamination. For instance, cells must be removed immediately from fluid samples. Proteins should not be left at room temperature and can be stored at 2–8°C for up to 3 d but at –20°C for longer periods. Thawing and refreezing should be minimized to avoid damaging the proteins. Vibration and pressure changes during transportation can also damage the proteins.
3. Consistency in the experimental protocol is of paramount importance. For instance, if too little sample is used, the small peaks will be below the limit of detection. Likewise, if too much sample is used, the peaks will be blunted, creating the same censoring problem as densitometer maladjustment.

3. State-of-the-Art Analytical Methods

We now review the analytical methods implemented in software packages such as MELANIE ***(18–20)*** and HERMeS ***(21–25)***. Let $I(x,y)$ denote the two-dimensional image, and, by convention, the larger $I(x,y)$ is, the darker the pixel is.

3.1. Filtering Gel Images

To reduce the high-frequency background noise, the signal is extracted by applying a smoothing filter. The most popular filters are

1. *Gaussian smoothing*, which convolves the image with the operator

$$\frac{1}{16}\begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix},$$

2. *diffusion smoothing*, that is,

$$\begin{aligned} I^{(t+1)}(x,y) &= \frac{1}{2}\left(I^{(t+1)}(x-1,y) + I^{(t)}(x+1,y)\right) \\ I^{(t+2)}(x,y) &= \frac{1}{2}\left(I^{(t+1)}(x-1,y) + I^{(t+2)}(x+1,y)\right) \\ I^{(t+3)}(x,y) &= \frac{1}{2}\left(I^{(t+3)}(x,y-1) + I^{(t+2)}(x,y+1)\right) \\ I^{(t+4)}(x,y) &= \frac{1}{2}\left(I^{(t+3)}(x,y-1) + I^{(t+4)}(x,y+1)\right), \end{aligned}$$

3. *polynomial smoothing*, where the pixel intensities in a small area (e.g., 3 × 3, 7 × 7) are approximated by a second-degree polynomial function in x and y

4. *adaptive smoothing*, that is,

$$I^{(t+1)}(x,y) = \frac{1}{N^{(t)}} \sum_{i=-1}^{1} \sum_{j=-1}^{1} I^{(t)}(x+i, y+j)\, w^{(t)}(x+i, y+j)$$

with

$$w^{(t)}(x,y) = \exp\left(-\frac{\left(d^{(t)}(x,y)\right)^2}{2K^2} \right), \quad N^{(t)} = \sum_{i=-1}^{1} \sum_{j=-1}^{1} w^{(t)}(x+i, y+j), \quad d^{(t)}(x,y) = \sqrt{G_x^2 + G_y^2}$$

where G_x and G_y are the gradients along the x- and y-axis, respectively ***(20)***.

Gaussian deconvolution is an alternative approach to remove noise and blur ***(2)***. Each spot is modeled as

$$I(x,y) = \sum_{k=-m}^{m} A(x+k, y)\; g_k + e(x,y) \text{ with } A(x+k,y) \geq 0 \text{ and } g_k = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-k^2}{2\sigma^2} \right)$$

where m is the integer part of 3σ and $e(x, y)$ represents random noise. Estimates are obtained via the constrained least-squares procedure. This approach tends to be overly sensitive to noise and to oversplit spots ***(26)***.

3.2. Spot Detection

For automatic spot detection, nonparametric procedures, based either on second derivatives ***(19,20)*** or on mathematical morphology ***(2)***, are utilized.

Let $\boldsymbol{p} = (x, y)$ be a point on the image, S_i a spot, and T the saturation threshold

$$\max(I) - \frac{100 - saturation}{100} \left(\max(I) - \min(I) \right)$$

where $0 \leq saturation \leq 100$ ($saturation = 100$ when no pixel is saturated), and $\Delta I(\boldsymbol{p})$ the Laplacian

$$-\left(\frac{\partial^2}{\partial x^2} I(\boldsymbol{p}) + \frac{\partial^2}{\partial y^2} I(\boldsymbol{p}) \right).$$

Is $\boldsymbol{p}$ part of the spot S_i? Select thresholds l, r, c (i.e., small positive constants). If $I(\boldsymbol{p}) < T$,

$$\boldsymbol{p} \in S_i \quad \Leftrightarrow \quad \min\left(\frac{\partial^2}{\partial x^2} I(\boldsymbol{p}) - r, \frac{\partial^2}{\partial y^2} I(\boldsymbol{p}) - c \right) > 0$$

when $-\Delta I(\boldsymbol{p}) - l \geq 0$. If $I(\boldsymbol{p}) > T$,

$$\boldsymbol{p} \in S_i \quad \Leftrightarrow \quad \min\left(\frac{\partial^2}{\partial x^2} I(\boldsymbol{p}),\; \frac{\partial^2}{\partial y^2} I(\boldsymbol{p}) \right) > 0 .$$

Small values of l allow the detection of as many spots as possible, while high values only yield dark spots with flat spots being ignored. High values of r and c help separate spots that are in close proximity of each other and to eliminate streaks. The algorithm identifies the spots by searching for the most negative values of the Laplacian and the two second derivatives (***19,20***). The Laplacian is indeed most negative at local peaks, while at the inflection points between unresolved spots the minimum value of the two second derivatives is negative.

With mathematical morphology, one can study characteristics of objects by investigating whether a standard shape fits into them. In this context, one searches for elevations relative to the local background brightness, and constructs an image based on the heights of these elevations. This is achieved by subtracting the *closing* of the image from the original image. This is the so-called *top-hat transform*, which in essence assesses whether a cylinder of a chosen radius fits into the elevations. To obtain the closing of the image, one first replaces each pixel value $I(x, y)$ with the local minimum intensity in a disk around each pixel, and then one replaces the resulting pixel value $I'(x, y)$ with a local maximum intensity, that is,

$$I''(x,y) = \max_{k,l} I'(x+k, y+l)$$

where $\sqrt{k^2 + l^2} \leq R$ (the disc radius) and $I'(x,y) = \min_{k,l} I(x+k, y+l)$.

In this closing, only pixels within elevations narrower than the chosen disc size have changed their values from the original image, and thus will show when the two images are subtracted. The radius R is selected to be the smallest value so that the disk is larger than the smallest spot ***(2)***. Shapes (or *structuring elements*) other than disks can be considered (e.g., spheres ***[27]***).

Alternatively, instead of a fixed structuring element, one can look for all h-domes ***(21,28,29)***. An h-dome is a connected region of pixels with intensity above h and greater than any pixel bordering the h-dome. These h-domes are not constrained by size. Algorithms for searching for these regions are more complex than the top-hat transform. The choice of h is crucial: if it is too small, background streaks will be recovered rather than spots, and, if it is too large, narrow peaks due to high-amplitude noise will be selected. Raising h stepwise allows the resolution of overlapping spots ***(28)***.

3.3. Background Filtering

The smooth background noise, which consists of vertical and horizontal streaks, is removed, either by subtracting the global minimum pixel value from all pixels or by estimating the background outside the spots with a third-order polynomial function ***(20)***. Because the background varies significantly across the image, a single threshold tends to work poorly: when it is set too high, faint spots are lost and, when it is set too low, high background is regarded as signal ***(28)***.

Mathematical morphology has also been utilized to remove the streaks *(21,27)*. One subtracts from the original image its closing with respect to two structuring elements, one vertical and one horizontal bar of one-pixel width, the lengths of which are slightly greater than those of the vertical and horizontal extents of the largest spots.

3.4. Spot Quantification

Spot characteristics are estimated by fitting two-dimensional Gaussian curves via the least-squares method:

$$g(x,y) = A \exp\left\{\left(\frac{x-x_c}{\sigma_X}\right)^2 + \left(\frac{y-y_c}{\sigma_Y}\right)^2\right\} + B$$

with A representing the amplitude, (x_c, y_c) the center, the σ's the spread along the principal axes and B the background level *(20,30,31)*. Spot models based on two half-Gaussian curves have also been utilized *(21)*. However, many spots are not Gaussian in nature because of several factors: local inhomogeneity within the acrylamide, overloading of the sample within the gel, adsorption of some proteins onto the acrylamide matrix, failure of some polypeptides to focus in the first dimension, and tendency for chemically distinct but barely resolved proteins to displace each other *(26)*.

Spot characteristics are thus better estimated directly:

$$\text{area} = \text{AREA} = \text{number of pixels} \times \text{pixel area}$$

$$\text{optical density} = \text{OD} = \max_{x,y \in \text{spot}} I(x,y)$$

$$\text{percent optical density} = \%\text{OD} = 100 \times \frac{\text{OD}}{\sum_{s=1}^{n} \text{OD}_s}$$

$$\text{volume} = \text{VOL} = \sum_{x,y \in \text{spot}} I(x,y)$$

$$\text{percent volume} = \%\text{VOL} = 100 \times \frac{\text{VOL}}{\sum_{s=1}^{n} \text{VOL}_s}$$

where OD_s and VOL_s are, respectively, the optical density and the volume of spot s in a gel containing n spots.

3.5. Image Alignment

Gels are aligned via polynomial image warping. Identify landmarks (or control points) on each image and choose one gel as the reference gel. The alignment algorithm attempts to superimpose these landmarks by stretching and shrinking

the images. Let (x,y) be the pixel coordinates in the original image and $[u(x), v(y)]$ those in the warped image. The latter are first-, second- or third-order univariable polynomials or their inverses. Estimate the parameters of these polynomial functions via the least-squares criterion by summing over the landmarks. Specifically, let there be M landmarks on each gel. The parameters are obtained by minimizing, for instance, if $M \geq 4$,

$$\sum_{i=1}^{M}\left(u_i - (a_0 + a_1 x_i + a_2 x_i^2 + a_3 x_i^3)\right)^2 \quad \text{and} \quad \sum_{i=1}^{M}\left(v_i - (b_0 + b_1 y_i + b_2 y_i^2 + b_3 y_i^3)\right)^2$$

where (u_i,v_i) refers to landmark i on the reference gel and (x_i,y_i) to its position on another gel. The value M determines the order of the polynomial: a polynomial of degree n requires at least $n + 1$ landmarks.

3.6. Spot Matching

Local gel-to-gel variations make it impossible to utilize a single transformation to map the spots from one gel to another. One approach is to divide the image into a number of small rectangular regions and to select, in each segment, 3 or more evenly spaced spots as reference points ***(28)***. These reference points serve to compute a transformation that maps spot centers from one film to another. Spots are considered matched if the transformed spot center from one gel and the corresponding spot center on the other gel are within 0.8 mm [a slightly more stringent criterion of 0.7 mm has also been utilized ***(32)***]. This procedure works best for the area defined by the reference points: spots located at the edges of the rectangular regions can be poorly matched. As a remedy, the following steps are added to the procedure: triangles of nearby matched spots are considered on both images, and the above algorithm is applied to the yet unmatched spots within these triangles.

Alternatively, for each spot on a gel, consider a cluster of neighboring spots ***(20)***. The central spot is regarded as the primary spot, and the surrounding spots as secondary spots. A spot belongs to a cluster if its centroid is inside a circle of fixed radius. This radius depends on the image dimension, number of spots on a gel, and minimum number of spots in the cluster. The clusters are characterized by polar coordinates centered at the primary spot. First, match the clusters with highest-intensity primary spots. Compare clusters via a probabilistic similarity measure. The probability that the next random hit falls within a cluster where $m - 1$ spots have been matched is given by

$$p_m = \frac{A_s - A_{m-1}}{A_c - A_{m-1}}$$

where A_s is the sum of the secondary areas in the cluster, A_c the total area within the boundary of the cluster, and A_{m-1} the total area of matched spots. If N stands for the number of spots in one cluster,

Prob(at least m spots are matched in N trials)

$$= \sum_{h=m}^{N} \binom{N}{h} \prod_{i=1}^{h} p_i \prod_{i=h+1}^{N} (1-p_i) \approx \sum_{h=m}^{N} \binom{N}{h} p_G^h (1-p_G)^{N-h} \quad \text{where } p_G = \left(\prod_{i=1}^{m} p_i \right)^{1/m}.$$

That spots be reliably matched is of paramount importance in the creation of representative images and in subsequent pattern-recognition analyses. Proceed with a consistency check for possible mismatching:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} t_x \\ t_y \end{pmatrix} + \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

by ensuring that $L = AD - BC \approx 1$ (rotation). For each primary cluster, estimate parameters from each set of 3 matched spots. When $L = 1.0 \pm 0.25$ and the rotation angle is ±10 degrees, the pairing is declared suitable. To project the remaining spots in the clusters, estimate the rotation parameters $A, \ldots, D$ by the least-squares method from good matchings in the two clusters.

Artificial-intelligence methodology has also been used to match spot lists ***(23)***. Because they are not based on geometrical considerations, they should be able to cope better with discontinuous gel distortions. Spot clusters are described via the angles and distances between any two spots in the cluster. Distances are divided into 3 classes and angles into 16 classes. Measurements are coded via their class identifiers. Spots are then matched via syntactic pattern-recognition techniques. Heuristic rules are imposed to limit the number of searches. Isolated spots tend to be problematic for this approach.

3.7. Creating Synthetic Gels

To obtain a master image from at least 3 pairwise matched gels, first select a reference gel ***(20)***. Check that the spots on the reference gel are well matched to spots on two other gels. These form triangles of matched spots, that is, the *starting groups*. Extend the starting groups by adding spots using the connectivity test: a spot must be matched with at least one other spot in the initial group. When all spots on the reference gel have been considered, create additional groups with the spots on the second gel that are not part of a group. Repeat with the other gels.

The synthetic gel contains the same number of spots as there are determined groups (these are the representative spots). The position of a spot on the synthetic gel is taken from the reference gel if the group has a spot on the reference gel. Otherwise, one translates the coordinates of the closest spot by considering a set of neighbors that have representatives on the reference gel. The intensity of the master spot is the average over the spots in the group. Its shape is the shape of the spot in the group that is closest in area to the average of the group ***(20)***.

To ensure the reliability of the master gel, some investigators compute an overlap measure between each master spot and the corresponding spots

(considered one at a time) on the aligned gels ***(1)***. This in effect assesses the quality of the matches on which the master gel relies. The overlap measure is simply the value of a Gaussian function evaluated at the physical distance of the spot centers. This Gaussian function is chosen to have height of 1.0 and width depending on the matching criterion: for instance, 0.7 mm if spots on different aligned gels are considered matched when they are within 0.7 mm ***(32)***.

3.8. Pattern Recognition

For pattern-recognition purposes, only spots that yield highly reliable features on the master image are considered in the analyses. For instance, the overlap measure between the master spot and a corresponding object on one of the aligned gels (cf. **Subheading 3.7.**) needs to be above a specific threshold (typically 0.5) and to exceed 90% of the largest overlap of the master spot with the original spots or of the overlap of the gel spot with all the master spots ***(1)***.

To find significant protein patterns associated with a specific disease, investigators have first recourse to principal-component analysis ***(1,25)*** or correspondence analysis and factor analysis to reduce dimensionality ***(18–20)***. This requires the computation of the normalized observation table so that the columns have mean 0 and variance 1. In factor analysis, the eigenvalues and eigenvectors of the covariance matrix are extracted to determine a factorial space (usually of dimension between 1 and 3). The gels are projected as points onto the factorial space. Spots can also be mapped onto this space so that characteristic spots can be identified: spots fall within the cluster of gels they typify.

These authors then apply clustering algorithms to the transformed data. The difficulty here is to define a meaningful distance metric. With principal-component analysis, one candidate is the Euclidean distance in the transformed space after weighing each coordinate by the percentage of the total variance represented by the corresponding principal component ***(1)***. The usual hierarchical clustering procedures, based on complete or single or average linkage, are utilized ***(1,18–20,33)***. A heuristic clustering algorithm has also been proposed ***(18,19)***. Suppose that n gels are to be classified into k classes. Select k gels by maximizing the Euclidean distance between them. These define k classes. A heuristic search is then performed: each of the remaining $n - k$ gels is included into one class and class descriptions are formulated. Iterate this process by choosing one gel per class, excluding the first k gels, to form new classes and repeating the previous step. This process continues until the classification converges.

One is in effect searching for protein patterns that best distinguish two groups of images (one from a “disease” group and one from a “control” group). Classification procedures would be best suited for this purpose. In actuality, patterns are identified, ignoring the associated outcomes, and then the inferred patterns are reconciled to the known groups.

4. Brief Introduction to Wavelets

Wavelets are building-block functions like *sine* and *cosine* functions in the Fourier transform ***(34)***. They oscillate about 0 and dampen to 0. This localization in time or space renders them highly versatile to model signals with nonsmooth features or that vary over time or space. The *father wavelet* or *scaling function* ϕ represents smooth, low-frequency components while the *mother wavelet* ψ represents detail, high-frequency components:

$$\int_{-\infty}^{+\infty}\phi(t) = 1 \quad \text{and} \quad \int_{-\infty}^{+\infty}\varphi(t) = 0 \,.$$

A number of orthogonal wavelet families have been constructed; for instance, Haar wavelets (symmetric square waves with compact support), daublets (continuous waves with compact support, d2 to d20 in S-plus ***[35]***), symmlets (nearly symmetric waves with compact support, s4 to s20 in S-plus ***[35]***).

Through a multiresolution analysis, one obtains fine to coarse resolution (scale) components of the signal, that is, for a one-dimensional signal,

$$f(t) = \sum_k s_{J,k}\phi_{J,k}(t) + \sum_k d_{J,k}\varphi_{J,k}(t) + \sum_k d_{J-1,k}\varphi_{J-1,k}(t) + \ldots + \sum_k d_{1,k}\varphi_{1,k}(t)$$

where J is the number of multiresolution components considered. The functions $\phi_{j,k}(t)$ and $\psi_{j,k}(t)$ are generated from ϕ and ψ by scaling and translation, that is,

$$\phi_{j,k}(t) = 2^{\frac{-j}{2}}\,\phi(2^{-j}t - k) = 2^{\frac{-j}{2}}\,\phi\left(\frac{t - 2^j k}{2^j}\right) \quad \text{and} \quad \varphi_{j,k}(t) = 2^{\frac{-j}{2}}\,\varphi(2^{-j}t - k) = 2^{\frac{-j}{2}}\,\varphi\left(\frac{t - 2^j k}{2^j}\right).$$

The scale/dilation factor 2^j affects the width of $\phi_{j,k}(t)$ and $\psi_{j,k}(t)$. The translation/location parameter $2^j k$ is coupled to the scale factor: as the support of $\phi_{j,k}(t)$ and $\psi_{j,k}(t)$ gets wider the translation steps become larger. As 2^j increases, $\phi_{j,k}(t)$ and $\psi_{j,k}(t)$ become shorter and more spread out. Finally, $s_{J,k}$, $d_{J,k}$, … , $d_{1,k}$ are the wavelet-transform coefficients:

scaling function coefficients $\quad s_{J,k} = \int_{-\infty}^{+\infty} f(t)\,\phi_{J,k}(t)\,dt$

wavelet coefficients $\quad d_{j,k} = \int_{-\infty}^{+\infty} f(t)\,\varphi_{j,k}(t)\,dt$

The $\phi_{j,k}(t)$'s and $\psi_{j,k}(t)$'s form an orthogonal basis:

$$\int_{-\infty}^{+\infty}\phi_{J,k}(t)\,\phi_{J,k'}(t)\,dt = \delta_{k,k'}\,, \quad \int_{-\infty}^{+\infty}\varphi_{J,k}(t)\,\phi_{J,k'}(t)\,dt = 0\,, \quad \int_{-\infty}^{+\infty}\varphi_{j,k}(t)\,\varphi_{j,k'}(t)\,dt = \delta_{k,k'}\delta_{j,}$$

where $\delta_{i,j} = 1$ if $i = j$ and 0 if $i \neq j$.

The discrete wavelet transform **W** for the discrete signal $\mathbf{f} = (f_1, f_2, \ldots, f_n)'$ is defined as

$$\mathbf{w} = \mathbf{W}\,\mathbf{f} \quad \text{where} \quad \mathbf{w}' = (\,\mathbf{s_J}'\,\mathbf{d_J}'\,\mathbf{d_{J-1}}'\,\ldots\,\mathbf{d_1}')$$

with
$\mathbf{s_J} = (s_{J,1}, s_{J,2}, \ldots, s_{J,n/2^J})'$, $\mathbf{d_J} = (d_{J,1}, d_{J,2}, \ldots, d_{J,n/2^J})'$, $\mathbf{d_{J-1}} = (d_{J-1,1}, d_{J-1,2}, \ldots, d_{J-1,n/2^{J-1}})'$, …,
$\mathbf{d_1} = (d_{1,1}, d_{1,2}, \ldots, d_{1,n/2})'$.

Each of the so-called crystals $\mathbf{s_J}$, $\mathbf{d_J}$, $\mathbf{d_{J-1}}$, **... ,** $\mathbf{d_1}$ contains the coefficients corresponding to a set of translated wavelet functions. In the multiresolution analysis,

$$f(t) \approx S_J(t) + D_J(t) + D_{J-1}(t) + \ldots + D_1(t),$$

the smooth and detail signals are represented, respectively, by

$$S_J(t) = \sum_k s_{J,k}\phi_{J,k}(t) \quad \text{and} \quad D_j(t) = \sum_k s_{j,k}\varphi_{j,k}(t) \quad j = 1, \ldots, J.$$

To compress an image, one utilizes a two-dimensional wavelet family

$$\Phi(x,y) = \phi_h(x) \times \phi_v(y) = \text{horizontal father} \times \text{vertical father}$$

$$\Psi^v(x,y) = \psi_h(x) \times \phi_v(y) = \text{horizontal mother} \times \text{vertical father}$$

$$\Psi^h(x,y) = \phi_h(x) \times \psi_v(y) = \text{horizontal father} \times \text{vertical mother}$$

$$\Psi^d(x,y) = \psi_h(x) \times \psi_v(y) = \text{horizontal mother} \times \text{vertical mother.}$$

The father wavelet Φ deals with the smooth aspect and the mother wavelets Y deal with the details in the vertical (Ψ^v), horizontal (Ψ^h), and diagonal (Ψ^d) dimensions. (**Figure 2** shows the diagonal s8 wavelet.)

The two-dimensional wavelet approximation is then

$$F(x,y) \approx \sum_{m,n} s_{J,m,n}\Phi_{J,m,n}(x,y) + \sum_{j=1}^{J}\sum_{m,n} v_{j,m,n}\Psi^v_{j,m,n}(x,y) + \sum_{j=1}^{J}\sum_{m,n} h_{j,m,n}\Psi^h_{j,m,n}(x,y) + \sum_{j=1}^{J}\sum_{m,n} d_{j,m,n}\Psi^d_{j,m,n}(x,y)$$

with

$$\Phi_{J,m,n}(x,y) = 2^{-J}\Phi(2^{-J}x - m, 2^{-J}y - n), \Psi^v_{j,m,n}(x,y) = 2^{-j}\Psi^v(2^{-j}x - m, 2^{-j}y - n),$$

$$\Psi^h_{j,m,n}(x,y) = \Psi^h(2^{-j}x - m, 2^{-j}y - n), \Psi^d_{j,m,n}(x,y) = \Psi^d(2^{-j}x - m, 2^{-j}y - n),$$

$$s_{J,m,n} = \int\int \Phi_{J,m,n}(x,y)F(x,y)\,dx\,dy, \; v_{j,m,n} = \int\int \Psi^v_{j,m,n}(x,y)F(x,y)\,dx\,dy,$$

$$h_{j,m,n} = \int\int \Psi^h_{j,m,n}(x,y)F(x,y)\,dx\,dy, \; d_{j,m,n} = \int\int \Psi^d_{j,m,n}(x,y)F(x,y)\,dx\,dy.$$

The two-dimensional discrete wavelet transform maps an $m \times n$ discrete image to $m \times n$ matrix of wavelet coefficients $\mathbf{w}_{m,n}$. In S-plus ***(35)***, $\mathbf{w}_{m,n}$ is decomposed into submatrices with coefficients for different multiresolution levels:

sJ – sJ	with coefficients $s_{J,m,n}$	for the smooth part
d1 – s1,..., dJ – sJ	with coefficients $v_{j,m,n}$	for the vertical detail

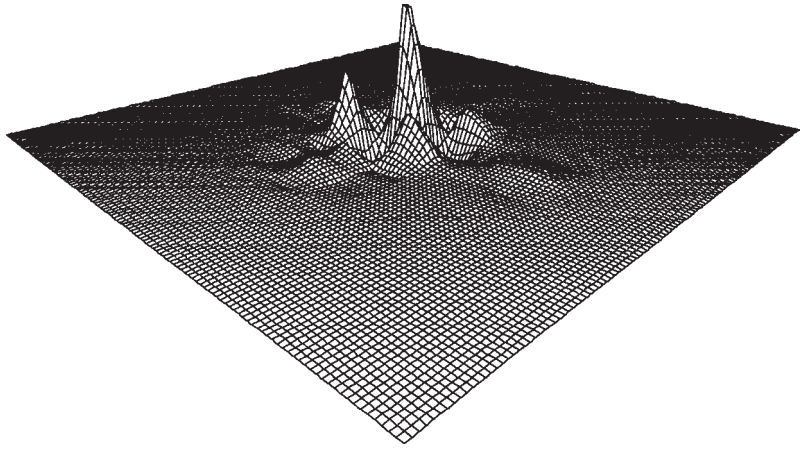

Fig. 2. s8 diagonal mother wavelet.

s1 – d1,..., sJ – dJ	with coefficients $h_{j,m,n}$	for the horizontal detail
d1 – d1,..., dJ – dJ	with coefficients $d_{j,m,n}$	for the diagonal detail.

Hence, in the multiresolution analysis,

$$F(x,y) \approx S_J(x,y) + \sum_{j=1}^{J} D_j^v(x,y) + \sum_{j=1}^{J} D_j^h(x,y) + \sum_{j=1}^{J} D_j^d(x,y)\,,$$

where

$$S_J(x,y) = \sum_{m,n} s_{J,m,n}\ \Phi_{m,n}(x,y) \quad , \quad D_j^h(x,y) = \sum_{m,n} h_{j,m,n}\ \Psi_{m,n}^h(x,y)\ ,$$

$$D_j^v(x,y) = \sum_{m,n} v_{j,m,n}\ \Psi_{m,n}^h(x,y) \quad , \quad D_j^d(x,y) = \sum_{m,n} d_{j,m,n}\ \Psi_{m,n}^d(x,y)\ .$$

5. Wavelets for Two-Dimensional Electrophoretic Data

Few statistical techniques can cope with the high dimensionality of 2D-PAGE data. Hence, for analytical reasons, the information is often reduced to the volumes of a manageable set of selected spots. This prohibits exploratory investigations of the data for the purpose of formulating testable hypotheses. We explored the possibility of fitting Gaussian curves. We selected a number of spots from the gel depicted in **Fig. 1** and assessed via statistical tests that we would not be justified in assuming that their shape is Gaussian (as is clearly evidenced by **Fig. 3**). We turned to wavelets for their versatility in representing irregular signals.

With this methodology, much effort is needed to identify the most suitable representation. Once the coefficients are selected, mainstream techniques can

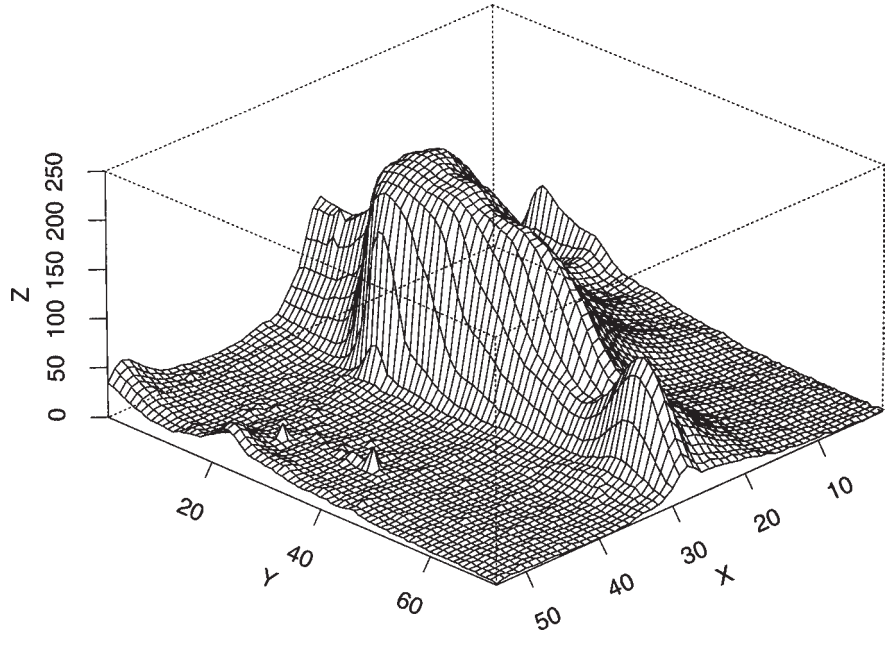

Fig. 3. Detail from kidney gel.

be applied to investigate the scientific questions of interest. We now review a few of these issues.

5.1. What Is the Most Suitable Wavelet Family to Represent Gels?

We considered Haar wavelets, daublets, and symmlets. All seemed suitable with the Haar family giving slightly worse results (**Figs. 4 – 7**). Even with an image corrupted by noise (Gaussian or Poisson), the reconstruction was highly successful (**Figs. 8 – 10**).

5.2. What Multiresolution Level Should One Select?

A high multiresolution level is not neccessary: most of the information is contained in the smoother crystals. The percentage of coefficients one wishes to retain is the more pertinent issue and depends on one's threshold for the volume of significant spots (**Figs. 11 – 14**).

5.3. How Does One Remove the Noise?

In the WaveShrink algorithm of S-plus *(35)*, one applies the wavelet transform with J levels then shrinks the detail coefficients

$$\tilde{\mathbf{d}}_1 = \delta_{\lambda_1\sigma_1}(\mathbf{d}_1), \ldots, \tilde{\mathbf{d}}_J = \delta_{\lambda_J\sigma_J}(\mathbf{d}_J)$$

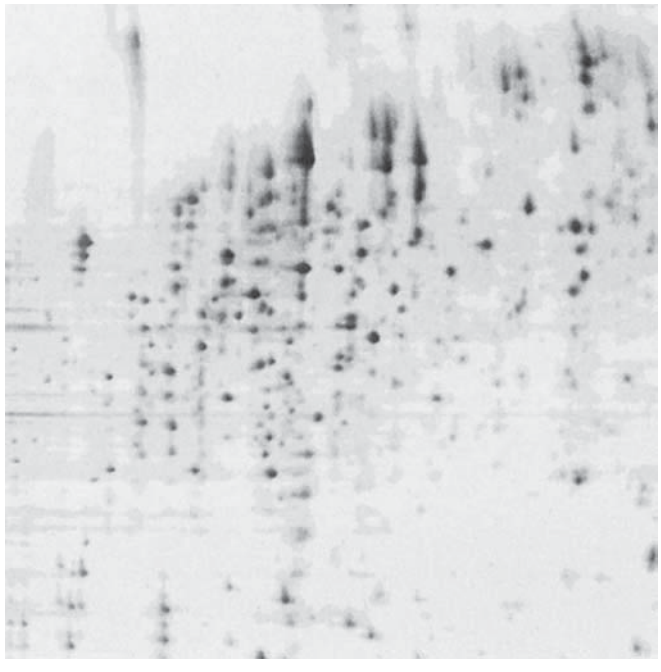

Fig. 4. Cropped image of kidney gel (512 × 512).

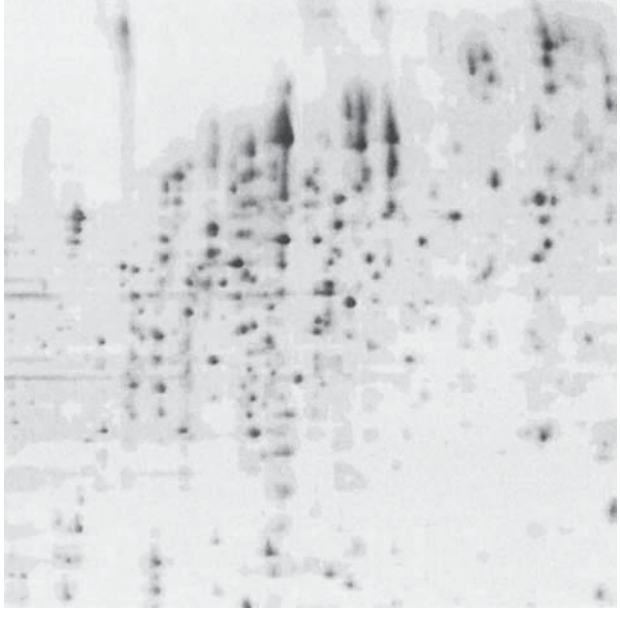

Fig. 5. Reconstruction of the cropped image of a kidney gel with daublets d4 at multiresolution level 4 and the largest 5% of the coefficients.

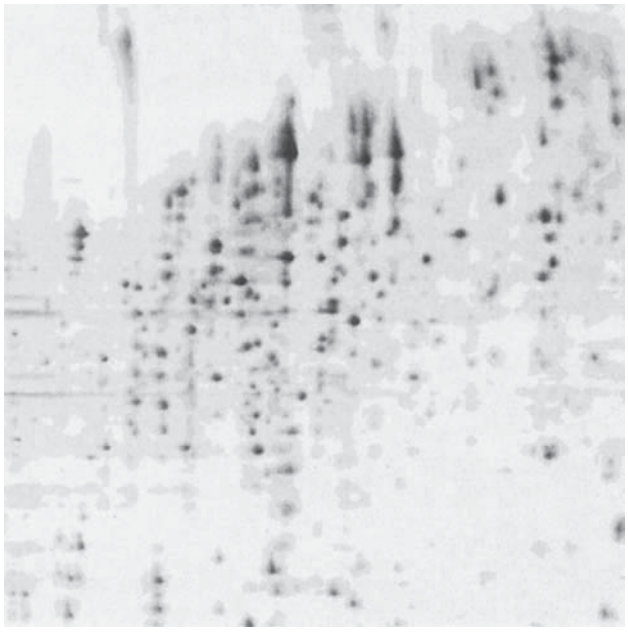

Fig. 6. Reconstruction of the cropped image of a kidney gel with symmlets s8 at multiresolution level 4 and the largest 5% of the coefficients.

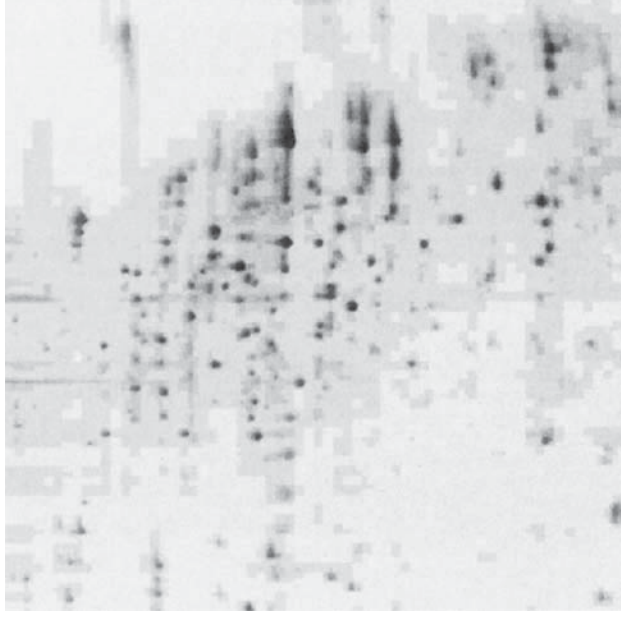

Fig. 7. Reconstruction of the cropped image of a kidney gel with Haar wavelets at multiresolution level 4 and the largest 5% of the coefficients.

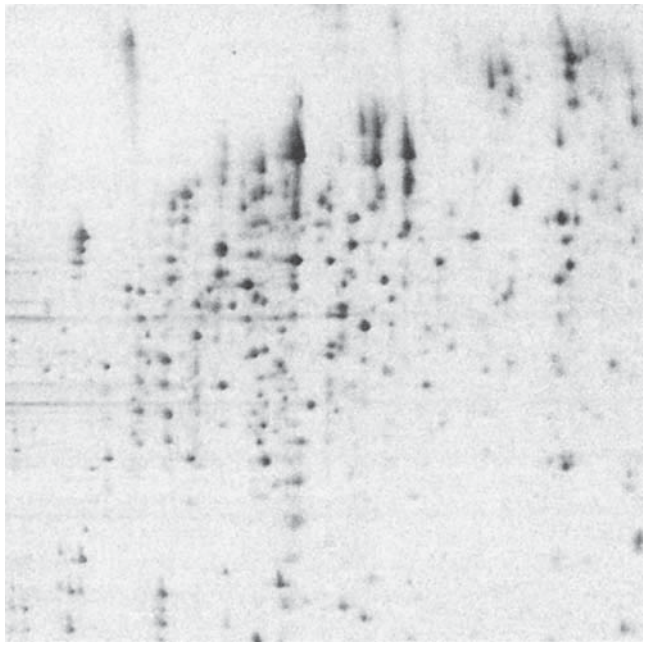

Fig. 8. Cropped gel image with added Gaussian noise (with variance 100).

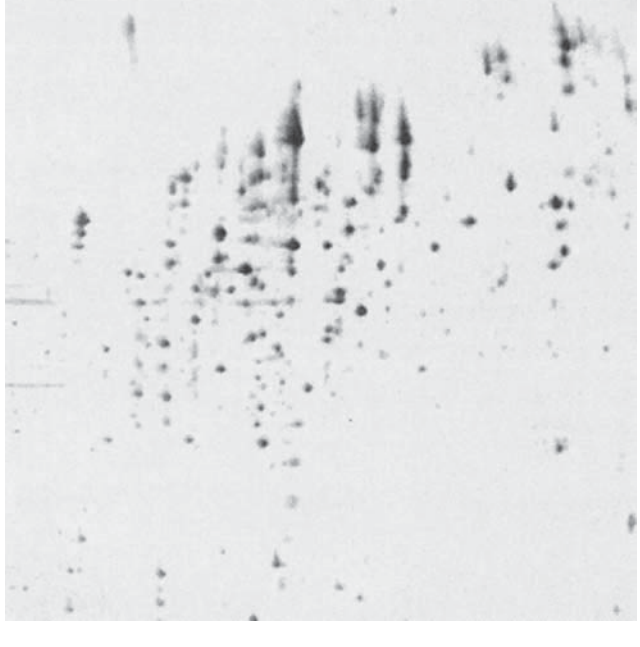

Fig. 9. Reconstruction of noisy gel with symmlets s8 at multiresolution level 1 and the largest 1% of the coefficients.

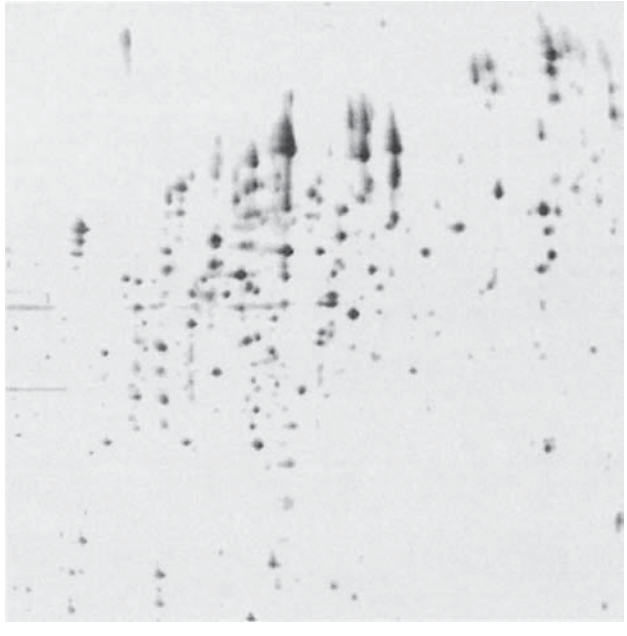

Fig. 10. Reconstruction of cropped image of kidney gel with symmlets s8 at multiresolution level 1 and the largest 5% of the coefficients.

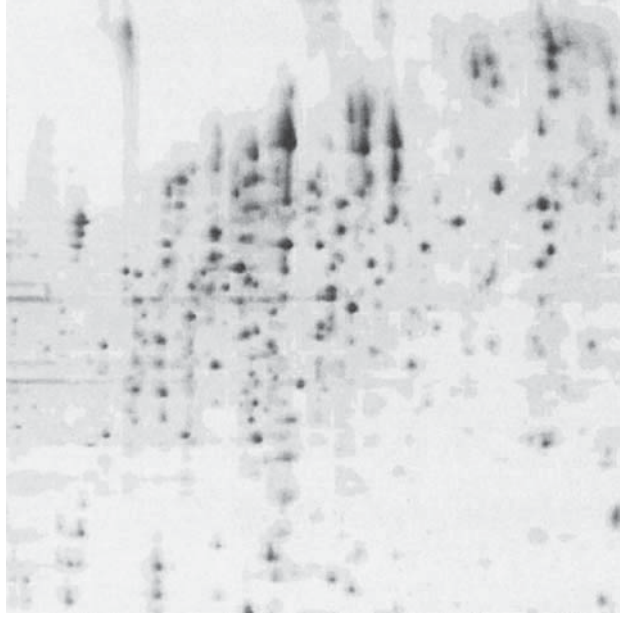

Fig. 11. Reconstruction of cropped image of kidney gel with daublets d4 at multiresolution level 3 and the largest 5% of the coefficients.

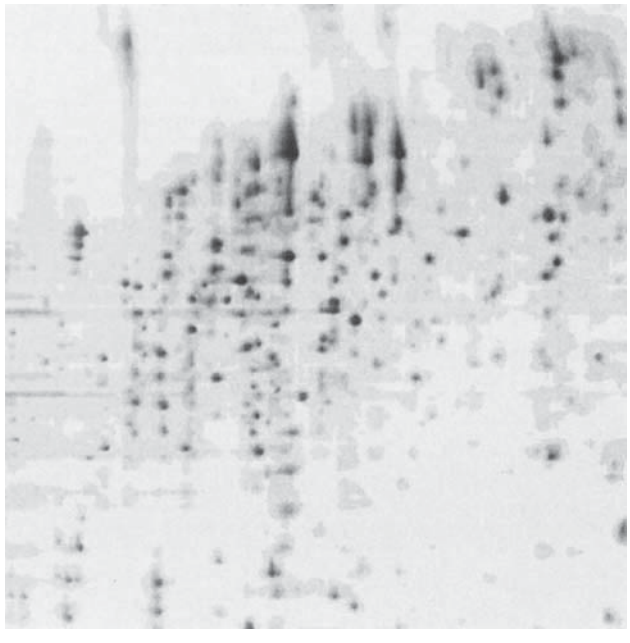

Fig. 12. Reconstruction of cropped image of kidney gel with daublets d4 at multiresolution level 5 and the largest 5% of the coefficients.

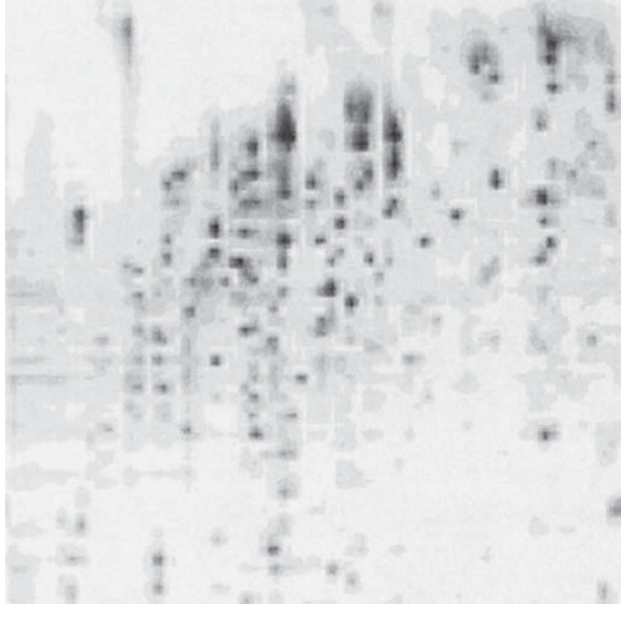

Fig. 13. Reconstruction of cropped image of kidney gel with daublets d4 at multiresolution level 3 and the largest 1% of the coefficients.

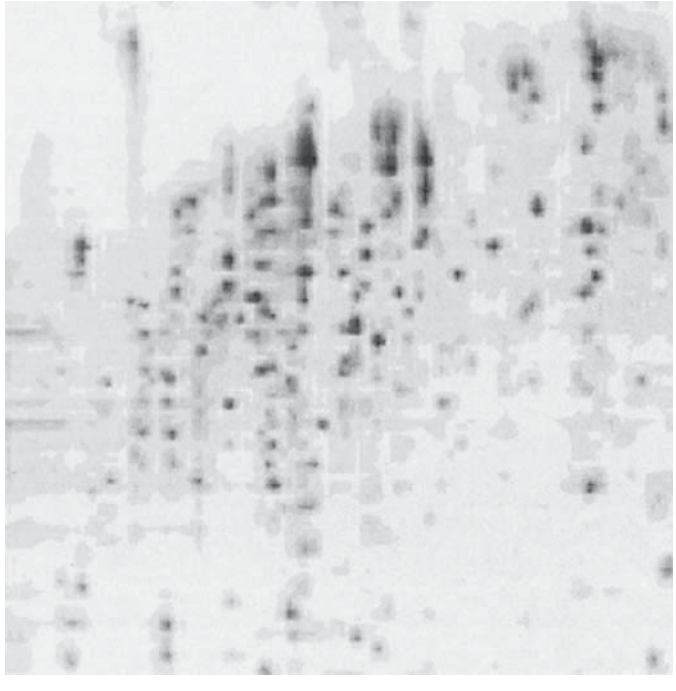

Fig. 14. Reconstruction of cropped image of kidney gel with daublets d4 at multiresolution level 5 and the largest 1% of the coefficients.

and reconstructs the image using $\tilde{\mathbf{d}}_1,\ldots,\tilde{\mathbf{d}}_J,\mathbf{s}_J$. Shrinkage is performed using the so-called soft or hard shrinkage functions

$$\delta_\gamma^S(x)=\begin{cases}0 & \text{if } |x|\le\gamma\\ \text{sign}(x)(|x|-\gamma) & \text{if } |x|>\gamma\end{cases},\quad \delta_\gamma^H(x)=\begin{cases}0 & \text{if } |x|\le\gamma\\ x & \text{if } |x|>\gamma\end{cases}.$$

For the thresholds λ_j, one can select the so-called universal value

$$\lambda_j=\sqrt{2\ \log n}$$

where n is the sample size. Alternatively, the value which minimizes the upper bound of the asymptotic risk (minimax) will result in less smoothing as it is always smaller than the universal threshold ***(36)***. Finally, we also considered Stein's unbiased risk estimator (SURE) which is adapted to each multiresolution level; the threshold for $\mathbf{d_j}$ with K coefficients is

$$\lambda_j=\arg\min_{t\ge 0}\text{SURE}(\mathbf{d_j},t)\ \text{ where }\ \text{SURE}(\mathbf{d_j},t)=K-2\sum_{k=1}^{K}1_{\{d_{j,k}\le t\sigma_j\}}+\sum_{k=1}^{K}\min\left((d_{j,k}\,\sigma_j)^2,\ t^2\right)$$

(37,38). To estimate the scale of the noise, that is, σ_j, one can rely either on the crystal corresponding to the finest detail or on all the detail crystals, or one can consider each crystal in turn, that is,

$$\tilde{\sigma}_j(\mathbf{d}_1),\ \ldots,\tilde{\sigma}_j(\mathbf{d}_1,\ldots,\mathbf{d}_J),\tilde{\sigma}_j(\mathbf{d}_j)$$

These algorithms yield rather poor results with the 2D-PAGE data (**Figs. 15–17**). Indeed, they are really smoothing techniques that are suitable to reduce highly localized and peaked noise. Here, the noise takes the form of streaks. All combinations we have tried result in the removal of important features.

We devised a hybrid procedure that seems to work well: hardshrinkage is utilized on the level 1 coefficients (these all tend to be very small) and the SJ - SJ crystal is multiplied by a constant between 0 and 1 (**Fig. 18**). This constant depends on the level of detail to be retained.

This routine is currently being optimized with respect to a biologically relevant objective criterion that involves the size of the spots being ignored.

5.4. How Does One Create a Master Gel?

Assume that the gels have been aligned. Wavelet coefficients are obtained for each of them. The synthetic gel is constructed by averaging the coefficients or by taking their median values. Variability in this construct can easily be computed.

5.5. How Does One Find Specific Protein Patterns for a Disease?

An analysis of the wavelet coefficients, for gels from diseased and control samples, based on classification and regression trees (CART), will highlight relevant clusters that best discriminate between the two groups. This has the advantage of considering both the location and the intensity of the spots simultaneously.

6. Conclusion

Electrophoresis has developed over the past 60 yr from a crude method able only to distinguish between very specific one-dimensional changes in experimental protocols to a highly complex technique. It is now possible not only to separate the genomic fingerprint of samples but also their proteome. While the technology has developed at an ever-increasing rate, the statistical techniques necessary to analyze such complex data structures has been left wanting. We have outlined some of the new methodologies that are currently available to take full advantage of the technology that is now in common usage in molecular biology laboratories.

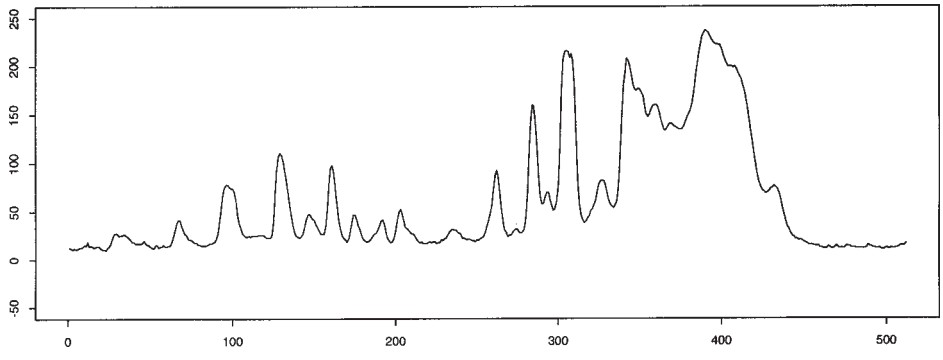

Fig. 15. Vertical slice of cropped image of kidney gel.

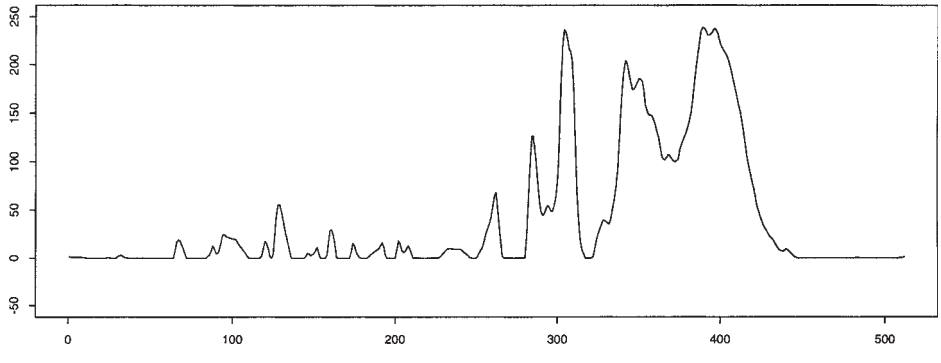

Fig. 16. Slice after hard shrinkage with universal threshold, estimating the scale of the noise separately for each crystal (multiresolution level 4).

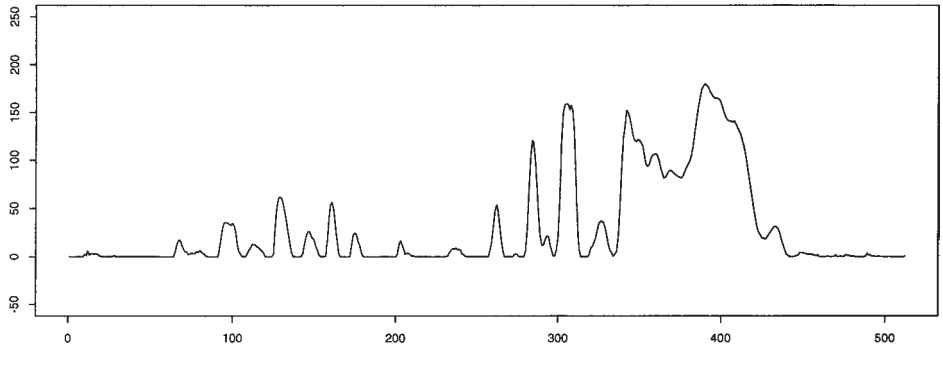

Fig. 17. Slice after soft shrinkage of the sJ - sJ crystal with universal threshold, estimating the scale of the noise from the sJ - sJ crystal (multiresolution level 4).

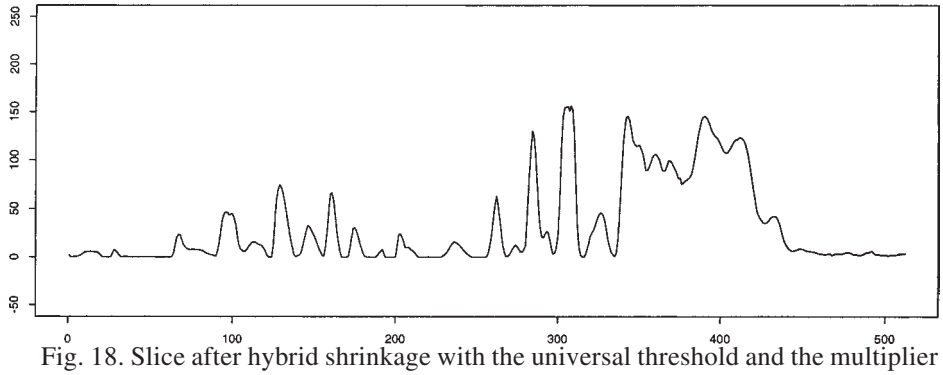

Fig. 18. Slice after hybrid shrinkage with the universal threshold and the multiplier set to 0.25 (multiresolution level 4).

References

1. Anderson, N. L., Hofmann, J. P., Gemmell, A., and Taylor, J. (1984) Global approaches to quantitative analysis of gene-expression patterns observed by use of two-dimensional gel electrophoresis. *Clin. Chem.* **30**, 2031–2036.
2. Horgan, G. W. and Glasbey, C. A. (1995) Uses of digital image analysis in electrophoresis. *Electrophoresis* **16**, 298–305.
3. Wilkins, M. R., Pasquali, C., Appel, R. D., Ou, K., Golaz, O., Sanchez, J-C., et al. (1996) From proteins to proteomes: large-scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Bio/Technology* **14**, 61–65.
4. Wilkins, M. R., Williams, K. L., Appel, R. D., and Hochstrasser, D. F. (eds.) (1997). *Proteome Research: New Frontiers in Functional Genomics.* Springer Verlag, New York.
5. Page, M. J., Amess, B., Rohlff, C., Stubberfield, C., and Parekh, R. (1999) Proteomics: a major new technology for the drug discovery process. *Drug Discovery Today* **4**, 55–62.
6. van Holde, K. E. (1985) Physical Biochemistry, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.
7. Tiselius, A. W. K. (1937) A new apparatus for electrophoretic analysis of colloidal mixtures. *Trans. Faraday Soc.* **33**, 524.
8. Burtis, C. A. and Ashwood, E. R. (1999) Tietz Textbook of Clinical Chemistry, 3rd ed. WB Saunders, Philadelphia.
9. Kenrick, K. G. and Margolis, J. (1970) Isoelectric focusing and gradient gel electrophoresis: a two-dimensional technique. *Analyt. Biochem.* **33**, 204–207.
10. Klose, J. (1975) Protein mapping by combined isoelectric focusing and electrophoresis in mouse tissues: a novel approach to testing for induced point mutations in mammals. *Humangenetik* **26**, 231–234.
11. O'Farrell, P. H. (1975) High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250**, 4007–4021.
12. Scheele, G. A. (1975) Two-dimensional gel analysis of soluble proteins: characterisations of guinea pig exocrine pancreatic proteins. *J. Biol. Chem.* **250**, 5375–5385.

13. Schägger, H. and von Jagow, G. (1987) Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 kDa. *Analyt. Biochem.* **166**, 368–379.
14. Wilkins, M. R., Sanchez, J.-C., Gooley, A. A., Appel, R. D., Humphrey-Smith, I., Hochstrasser, D. F., and Williams, K. L.(1995) Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. Biotechnol. *Genet. Engng. Rev.* **13**, 19–50.
15. Bjellqvist, B., Ek, K., Righetti, P. G., Gianazza, E., Görg, A., Westermeier, R., and Postel, W. (1982) Isoelectric focusing in immobilized pH gradients: principle, methodology and some applications. *J. Biochem. Biophys. Methods* **6**, 317–339.
16. Hochstrasser, D. F., Frutiger, S., Paquet, N., Bairoch, A., Ravier, F., Pasquali, C., et al. (1992) Human liver protein map: A reference database established by microsequencing and gel comparison. *Electrophoresis* **13**, 992–1001.
17. Laemmli, U. K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **277**, 680–685.
18. Appel, R., Hochstrasser, D. F., Roch, C., Funk, M., Muller, A. F., and Pellegrini, C. (1988) Automatic classification of two-dimensional gel electrophoresis pictures by heuristic clustering analysis: a step toward machine learning. *Electrophoresis* **9**, 136–142.
19. Appel, R., Hochstrasser, D. F., Funk, M., Vargas, J. R., Pellegrini, C., Muller, A. F., and Scherrer, J. R. (1991) The MELANIE project: from a biopsy to automatic protein map interpretation by computer. *Electrophoresis* **12**, 722–735.
20. Appel, R., Palagi, P. M., Walther, D., Vargas, J. R., Sanchez, J. C., Ravier, F., et al. (1997) MELANIE II — A third-generation software package for analysis of two-dimensional electrophoresis images: I. Features and user interface. *Electrophoresis* **18**, 2724–2734.
21. Vincens, P. (1986) HERMeS: a second generation approach to the automatic analysis of two-dimensional electrophoresis gels. Part II: Spot detection and integration. *Electrophoresis* **7**, 357–367.
22. Vincens, P., Paris, N., Pujol, J. L., Gaboriaud, C., Rabilloud, T., Pennetier, J. L., et al. (1986) HERMeS: a second generation approach to the automatic analysis of two-dimensional electrophoresis gels. Part I: Data acquisition. *Electrophoresis* **7**, 347–356.
23. Vincens, P. and Tarroux, P. (1987) HERMeS: a second generation approach to the automatic analysis of two-dimensional electrophoresis gels. Part III: Spot list matching. *Electrophoresis* **8**, 100–107.
24. Vincens, P. and Tarroux, P. (1987) HERMeS: a second generation approach to the automatic analysis of two-dimensional electrophoresis gels. Part IV: Data base organization and management. *Electrophoresis* **8**, 173–186.
25. Tarroux, P., Vincens, P., and Rabilloud, T. (1987) HERMeS: a second generation approach to the automatic analysis of two-dimensional electrophoresis gels. Part V: Data analysis. *Electrophoresis* **8**, 187–199.
26. Miller, M. J., Vo, P. K., Nielsen, C., Geiduschek, E. P., and Xuong, N. H. (1982) Computer analysis of two-dimensional gels: semi-automatic matching. *Clin. Chem.* **28**, 867–875.

27. Skolnick, M. M., Sternberg, S. R., and Neel, J. V. (1982) Computer programs for adapting two-dimensional gels to the study of mutation. *Clin. Chem.* **28**, 969–978.
28. Vo, K. P., Miller, M. J., Geiduschek, E. P., Nielsen, C., Olson, A., and Xuong, N. H. (1981) Computer analysis of two-dimensional gels. *Analyt. Biochem.* **112**, 258–271.
29. Vincens, P. (1993) Morphological grayscale reconstruction in image analysis. IEEE Trans. *Image Proc.* **2**, 176–201.
30. Lutin, K. W. A., Kyle, C. F., and Freeman, J. A. (1978) Quantitation of brain proteins by computer-analyzed two dimensions electrophoresis, in *Electrophoresis '78* (Catsimpoolas, ed.), *Developments in Biochemistry*, vol. 2, Elsevier, NY, pp. 93–106.
31. Garrels, J. (1979) Two-dimensional gel electrophoresis and computer analysis of proteins synthesized by clonal cell lines. *J. Biol. Chem.* **254**, 7961–7977.
32. Taylor, J., Anderson, N. L., and Anderson, N. G. (1981) A computerized system for matching and stretching two-dimensional gel patterns represented by parameter lists, in *Electrophoresis '81* (Allen, R. A. and Arnoud, P., eds.), W de Gruyter, NY, pp. 383–400.
33. Tarroux, P. (1983) Analysis of protein patterns during differentiation using 2-D electrophoresis and computer multidimensional classification. *Electrophoresis* **4**, 63–70.
34. Daubechies, I. (1992) Ten Lectures on Wavelets. Society for Industrial and Applied Mathematics, Philadelphia, PA.
35. S-Plus (2000) Data Analysis Products Division, MathSoft, Seattle, WA.
36. Donoho, D. L. and Johnstone, I. M. (1994) Ideal spatial adaptation via wavelet shrinkage. *Biometrika* **81**, 425–455.
37. Donoho, D. L. and Johnstone, I. M. (1995) Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Assoc.* **90**, 1200–1224.
38. Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995) Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B* **57**, 301–369.