Review

Andrew W. Dowsey[1]
Michael J. Dunn[2]
Guang-Zhong Yang[1]

[1]Royal Society/Wolfson Foundation Medical Image Computing Laboratory, Imperial College London, London, UK
[2]Department of Neuroscience, Institute of Psychiatry, King's College, London, UK

The role of bioinformatics in two-dimensional gel electrophoresis

Over the last two decades, two-dimensional electrophoresis (2-DE) gel has established itself as the *de facto* approach to separating proteins from cell and tissue samples. Due to the sheer volume of data and its experimental geometric and expression uncertainties, quantitative analysis of these data with image processing and modelling has become an actively pursued research topic. The results of these analyses include accurate protein quantification, isoelectric point and relative molecular mass estimation, and the detection of differential expression between samples run on different gels. Systematic errors such as current leakage and regional expression inhomogeneities are corrected for, followed by each protein spot in the gel being segmented and modelled for quantification. To assess differential expression of protein spots in different samples run on a series of two-dimensional gels, a number of image registration techniques for correcting geometric distortion have been proposed. This paper provides a comprehensive review of the computation techniques used in the analysis of 2-DE gels, together with a discussion of current and future trends in large scale analysis. We examine the pitfalls of existing techniques and highlight some of the key areas that need to be developed in the coming years, especially those related to statistical approaches based on multiple gel runs and image mining techniques through the use of parallel processing based on cluster computing and the grid technology.

Keywords: Bioinformatics / Data mining / Review / Spot detection / Spot matching / Two-dimensional gel electrophoresis PRO 0459

Contents

Correspondence: Dr. Guang-Zhong Yang, Department of Computing, 180 Queen's Gate, Imperial College, London SW7 2BZ, UK
E-mail: gzy@doc.ic.ac.uk
Fax: +44-20-7581-8024

Abbreviations: DN, Delavnay net; **IOD**, integrated optical density; **MIR**, Multiresolution Image Registration; **PCA**, principle component analysis; **SVD**, singular value decomposition

1 Introduction

Current development in genomics has provided a vast amount of information linking gene activity with disease [1]. It is now recognized, however, that there are a number of reasons why gene sequence information does not provide a complete profile of a protein's abundance or its final structure and state of activity. It has been estimated that for mammalian genomes a single gene can encode on average as many as six different protein species [2]. The proteome is therefore far more complex than the genome.

Since it is proteins that are directly involved in both normal and disease-associated biochemical processes, a more complete understanding of disease may be gained by looking at the proteins present within a diseased cell or tissue. This forms the basis of proteomics. The potential biological and clinical applications of proteomics are enormous [3].

The first stage of proteomics is sample collection. Samples are then pretreated by solubilization, denaturation and reduction to completely break the interactions between proteins and to remove nonprotein components. The next step is to isolate all the proteins from each other so they can be identified and quantified individually.

Two-dimensional polyacrylamide gel electrophoresis (2-DE) is the only method currently available which is capable of simultaneously separating and quantitating 10 000 proteins [4], and has dominated the field for more than 20 years [5]. While recent advances in quantitative mass spectrometry, particularly based on the use of stable isotope tags, are showing great promise [6], 2-DE remains the method of choice for the majority of studies of differential protein expression. The first dimension of 2-DE is isoelectric focusing, during which proteins are separated in a pH gradient until they reach a stationary position where their net charge is zero. The pH at which a protein has zero net charge is called its isoelectric point (p*I*). In the second dimension the proteins are separated orthogonally by electrophoresis in the presence of SDS according to their relative mass (M_r). Silver staining is considered the gold standard for detecting minor proteins [7], but other stains such as Coomassie Brilliant Blue (CBB) and SYPRO Ruby (Bio-Rad, Hercules, CA, USA) are also used. The gels can be transferred to computer for analysis using a high quality 12- to 16-bit greyscale scanner.

One of the key objectives of biochemistry is to identify the differential expression between control and experimental samples run on a series of 2-D gels. That is, the protein spots that have been inhibited (disappeared), induced (appeared) or have changed abundance (increased or decreased in size and intensity). Once these gel features are found, the proteins within them can be identified sensitively and accurately using MS.

Although at first glance the resolution of 2-DE seems impressive, it is still not sufficient compared to the enormous diversity of cellular proteins, and comigrating proteins in the same spot are not uncommon [8]. Neighboring spots can obscure proteins spot centers in these so-called complex regions, and their saturated nature can make the resolution of each individual protein intractable (Fig. 1(a)). Narrow pH range ("zoom") gels will reduce this problem but add a new challenge of sewing this patchwork quilt together [9]. Spots tend to have symmetric diffusion in the p*I* dimension but often severe tails in the M_r dimension. The diffusion depends on the protein concentration, which is why streaks and smears occur with certain proteins, as shown in Fig. 1(b).

There is an enormous divergent expression of proteins in cells and tissues. It has been estimated that the strongest third of spots account for more than 75% of the total amount of protein in the sample and the weakest third of the spots account for less than 6% of the total protein amount. The dynamic range between the least expressed and most expressed (amount of molecules present) proteins can be up to 10^6 for cells and tissues and as much as 10^{12} in body fluids such as plasma [10]! Whilst 2-DE has a maximum dynamic range of 10^4, at this value the scarcest proteins require an expert eye to discriminate valid spots from noise *e.g.* the faint spots in Fig. 1(c). Also, the intensity of the image background can vary across the image. Intensity profiles show a larger background variation in the vertical direction and higher background intensity at the edges of the gel than in the gel center. These artefacts can also be caused by the stain binding to nonprotein elements, *e.g.* silver stain binds to DNA and lipopolysaccharide [7].

These are some of the challenges facing the automation of spot detection in the bioinformatics pipeline. But why do we need to apply statistical and computational techniques to proteomics? For a biochemist to analyze a pair of gels for differential expression, bearing in mind the thousands of candidate proteins, would require several hours of expert analysis. The process quickly becomes impossible when we notice the analogue deformations in the electrophoretic diffusion process make it very hard to even match each protein spot on one gel to the same protein spot in the other. Staining variation between gels can cause weak spots to have invisible partner spots in the other gel.

The geometric distortions of the protein patterns are due to the casting, polymerization and running procedure of the gels. Four factors have been identified: (i) the structure of the media (the polyacrylamide net); (ii) the characteristics of the transporting solute; (iii) the solvent conditions (buffer); and (iv) the nature of the electric field.

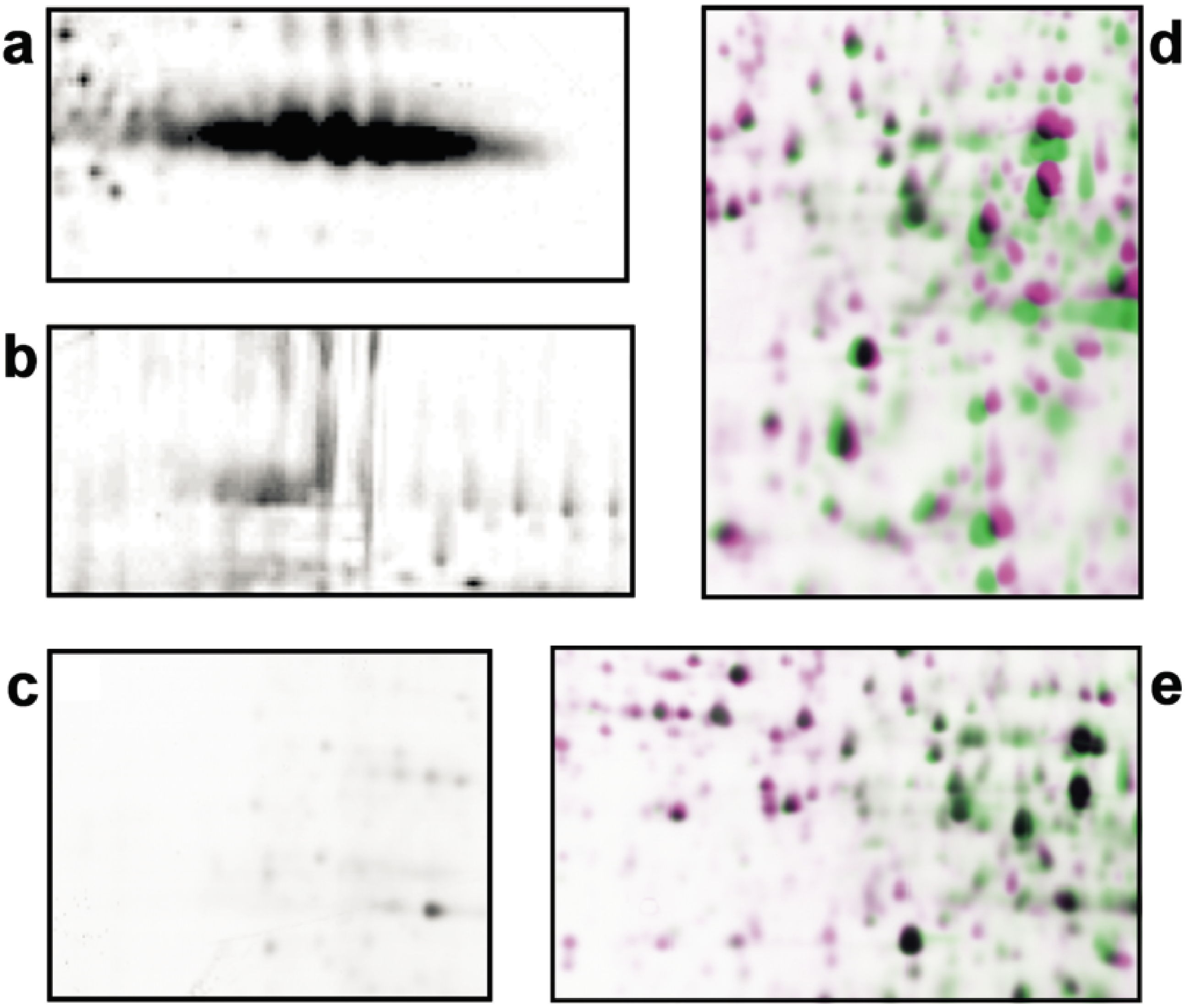

Figure 1. Common problems for computational image analysis as illustrated on silver stained gels: (a) comigrating spots forming a complex region; (b) streaking and smearing in the electrophoresis dimension for some proteins and protein concentrations, (c) weak spots, and background inhomogeneity caused by stain binding to nonprotein elements. Two superimposed gels, one green and one magenta, show (d) nonlinear local distortions between the gels; (e) and after spatial registration, regional intensity inhomogeneities which prove a source of systematic error for spot quantification.

There are two dominant models of gel migration, the standard Ogston-Morris-Rodbard-Chrambach model [11], and the reptation model [12] (as in reptile-like motion). The most important factor is due to current leakage, a global change of the electric field. However, there are still many more local distortions present due to combinations of minor factors, as is shown by the variable displacements between corresponding proteins in Fig. 1(d). For instance, fixing the gel can cause it to shrink and swell unevenly.

In difference gel electrophoresis (commercially available as Ettan DIGE; Amersham Biosciences, Uppsala, Sweden) [13] succinimidyl esters of the cyanine dyes are employed to fluorescently label two different complex protein mixtures prior to mixing them together and running them simultaneously on the same 2-D gel. The gel images are then acquired using two different emission filters. DIGE removes the requirement for matching these intra-gel samples, though of course it does not affect the requirement for inter-gel samples. Recently a novel approach to this problem has been suggested in which a pooled "standard" sample labelled with a third dye is included in each experimental pair of labelled samples run on each 2-D gel. This pooled standard sample is then used to normalize protein abundance measurements across multiple gels in an experiment [14, 15] and a dedicated software platform, DeCyder (Amersham Biosciences) has been developed for such analysis. Nevertheless, this approach is still dependent on the accurate matching and comparison of large sets of 2-D gel images in order to generate meaningful data on differential protein expression between sets of samples.

The silver staining technique is well known to be far from stoichiometric [16]. The intensity is only linear over a 40- to 50-fold low nanogram range in concentration (CBB has a 20-fold high nanogram range and SYPRO Ruby has a 1000-fold range [7]). Above this the stain density becomes nonlinear as spot densities reach saturation. Silver staining density and protein concentration is also dependent on the type of protein – the amino acid concentration and post-translational modifications. Contrast varies from gel to gel due to stain exposure [17], sample loading errors and protein losses at different stages of gel running, which is a challenge in quantifying differential expression. In Fig. 1(e) expression in one gel becomes weaker from left to right as it strengthens on the other. Important changes in protein expression may be obscured using only two gels, so that multiple experimental runs of the same samples could be carried out. Integrating these multiple runs would require a new rigorous statistical approach model. Protein quantification and differential expression would be computed as probabilities and illustrated as levels of confidence, rather than false positive/negative errors propagating down the pipeline.

Given all these problems it becomes very arduous, repetitive and time-consuming for a biochemist to identify and quantify patterns of differential protein expression in his experiments. There are great economic and efficiency reasons to eliminate this bottleneck with the use of computation, however these same problems pose a great challenge for fully automated gel processing [5]. If the goal of the experiment is to look for quantitative changes in the biological process or look quantitatively at protein modifications, then 2-DE will remain unrivalled for some time [5]. Therefore, the realization of large-scale proteomics for drug discovery and proteome mapping, requiring throughput of thousands of samples each day, will need a solution for full automation of the image analysis pipeline.

1.1 Image acquisition

The first step in computerized image analysis of 2-D gel protein profiles is capture of the gel images in a digital format. A range of devices, including modified document scanners, laser densitometers, charge-coupled device (CCD) cameras, and fluorescent and phosphor imagers, are available for the acquisition of 2-D gel images. Although most analysis systems can interface with most image digitization devices, selection of the appropriate device depends largely on the types of detection systems used in a particular laboratory to visualize 2-D gel separations. CCD cameras and densitometers based on "enhanced" document scanners are good general purpose devices as they permit analysis of gels visualized with a variety of stains. However, these devices are more susceptible to greyscale saturation effects than other detectors, such as laser densitometers. In addition, CCD cameras and document scanners are usually 8 bit devices with a maximum optical density (OD) range of no more than 200 to 1. In contrast, laser densitometers, and fluorescent and phosphor imagers are usually 12-to 16-bit devices with response ranges up to 10^5 to 1. Clearly, fluorescent imaging devices are restricted to use with fluorescent dyes (*e.g.* DIGE, SYPRO Ruby), while phosphor imagers are used in preference to densitometry of X-ray film with radiolabelled samples. See Miller *et al.* [18] and Miura [19] for a comprehensive review of image acquisition techniques.

1.2 Packages

The traditional pipeline for a 2-DE software package is [20]: (a) Pre-processing of the gel images – image normalization, cropping and background subtraction; (b) Spot segmentation (detection) and expression quantification; (c) An initial user guided pairing of a few spots between the reference and sample gels (landmarking). The sample gel is then warped to align the landmarks; (d) An automatic pairing of the rest of the spots; (e) Identification of differential expression; (f) Data presentation and interpretation; and (g) Creation of 2-D gel databases.

Most packages offer a host of auxiliary features such as: import and export of images in popular formats (TIFF, PNG, BMP *etc.*); annotating spots; querying spot lists; and connection to online databases. A good proportion of packages on the market today are based in part to the original academic developments of the late 1970's and early 1980's. For example TYCHO [21], GELLAB [22–25], HERMeS [26, 27], QUEST [28], LIPS [29], GESA [30] and ELSIE [31]. Some of these form the base of commercial packages, which include: ([32] provides a good list) Delta 2D (DECODON, Greifswald, Germany); PDQuest (Bio-Rad, Hercules, CA, USA); Phoretix 2D and Progenesis (Nonlinear Dynamics, Newcastle upon Tyne, UK); α-GelFox 2D (Alpha-Innotech, San Leandro, CA, USA); Image Master 2D and DeCyder (Amersham Biosciences); GELLAB-II (Scanalytics, Fairfax, VA, USA); Melanie 3 (GeneBio, Geneva, Switzerland); Investigator HT Analyzer (Genomic Solutions, Ann Arbor, MI, USA); KEPLER (Large Scale Biology, Germantown, MA, USA); and Bio Image 2D Investigator (Genomic Solutions). All of these are now closed source. However, the image analysis problems in 2-DE are similar to those in other fields, such as medical imaging and geoscience, so we do not expect these packages to stray too far in their implementation. Papers with varying levels of detail are available for CAROL [32], Melanie [33, 34], GELLAB [22–25], PiKA^2 [35–37], TEX [38], Flicker [39, 40] and Z3 [41].

1.3 Precision and reliability of existing systems

In 1989 Miller and Merril [42] constructed a series of tests to analyze the precision, reliability and reproducibility of gel analysis systems. The output stability of the acquisition device is tested. Spot resolution and quantification is tested with the aid of scatter plots on synthetic gels, with and without generated noise, and real gels. Synthetic gels have the advantage that they can be tailored but they do not truly represent the real world situation. They found ELSIE 4 spot quantification varied by 13% between different exposures and 15% between different gels run with the same sample.

Mahon and Dupree [43] conducted experiments to deduce the reliability of Phoretix 2D in determining changes in individual protein levels in complex protein mixtures with quantitative 2-DE. They compared the volume of segmented protein spots between multiple scans of the same gel and between multiple gels of the same experiment. Errors were categorized into spot matching errors and volume quantization errors. For the scan-to-scan experiment, relative spot matching errors neared 0% and relative quantization errors fell between 1%–10% depending on the decreasing abundance of the protein. However, between gels spot matching errors increased to 10% and notably, quantization errors were five to tenfold larger.

Nishihara and Champion [7] compared Z3, Progenesis and PDQuest, and found all to have a coefficient of variation in their spot detection reproducibility of 4–11% for SYPRO Ruby staining. Spot quantification reproducibility lay between 3–33%, where the higher values were for proteins of lower abundance. Using Z3 to detect differential expression they found that with a threshold of '2' still 10% of the result was misidentified by staining inhomogeneities. Raman *et al.* [32] developed tests for comparing spot detection, matching and quantization between packages, and contrasted Z3 and Melanie. Spot detection was carried out on real gels against manual detection, matching was carried out on the same real gel artificially distorted, and quantification was carried out on artificial gels of Gaussian spots. In the spot detection Z3 had 11% false negatives and 14% false positives, and Melanie more. Spot matching errors lay between 3–10%. Melanie fared significantly better than Z3 in spot quantification. Mahon and Dupree [43] suggest that these errors are a consequence of pipetting, gel focusing and staining errors. Also, since 76% of the variation in volume in the scan-to-scan data can be attributed to errors in spot area, there is strong evidence that the spot segmentation process effects the protein quantization significantly.

Recently Rogers *et al.* [44] formulated an objective method of comparing spot detection algorithms by significantly improving the creation of 'ground truth' synthetic gels. By using a new statistical spot model [45] (see Section 2.2) trained from a set of real gels, together with noise, they are able to create realistic images with unambiguous interpretation. The sensitivity of the spot detection parameters in ImageMaster, Melanie III, PDQuest, Progenesis and Z3 were illustrated on free response operator characteristic (FROC) curves. Ten images with varying S/N were used to evaluate their tolerance to noise and 40 images were used to evaluate sensitivity to spot overlap. ImageMaster was found to be the most accurate package with 85.1% true positives. Z3 coped the best with noise and PDQuest the most robust to spot overlap. Melanie III performs consistently well in all areas and Progenesis had the advantage of parameter free spot detection. The accuracies of these results show that we are still a long way from a totally automatic gel processing system that does not require user intervention. For proteins of low abundance a major factor is the sensitivity of current algorithms to noise, and for saturated proteins the errors are systematic to the electrophoresis process.

2 Singleton image analysis

This discipline concerns the image processing of isolated gels to correct for systematic experimental errors with the aim of accurately segmenting proteins to measure the pH and molecular mass of each protein and accurately quantifying their expression. The image processing pipeline traditionally includes: (a) Pre-processing for streak/noise removal, background correction and intensity normalization; (b) spot segmentation and modelling; and (c) quantification of each spot.

2.1 Image correction

Before examining the entire pipeline as a whole, it is necessary to discuss standard gel preprocessing techniques and correction of the systematic gel inhomogeneities caused by the 2-DE process for more accurate information extraction.

2.1.1 Spatial correction

Gustafsson *et al.* [46] assume the major factor in the geometric distortions is the current leakage due to a global change in the electric field. Correcting this leads to simpler p*I* and M_r measurements plus the registration phase only has to compensate for the local distortions of the other minor factors. Previous methods [20] defined a M_r/p*I* grid where p*I* is linearly interpolated and M_r exponentially interpolated from the locations of a set of known proteins (markers) on the gel.

Current leakage in the electrophoresis phase causes the global 'frown' in gels whose sides are not fully isolated. Electrostatic potential in the gel Ψ is given by Poisson's equation (not given), but assuming electro-neutrality (net charge volume density = 0) and uniform permittivity, it reduces to Laplace's equation:

$$\nabla^2\Psi = 0 \tag{1}$$

with Ψ bounded by the applied voltage $-V_0$ at the start (gel top – cathode) and 0 at the finish (gel bottom – anode). The Ψ boundary conditions at the gel sides are:

$$\begin{aligned}\frac{\partial\Psi}{\partial n} + \gamma_L\Psi = 0 \quad &\text{at the left boundary}\\ \frac{\partial\Psi}{\partial n} + \gamma_R\Psi = 0 \quad &\text{at the right boundary}\end{aligned} \tag{2}$$

where $\partial\Psi/\partial n$ is the outward normal derivative of Ψ at the boundary and the γ are the current leakages:

$$\gamma = \frac{\alpha_{\text{spacer}}}{\alpha_{\text{gel}}\Delta w} \tag{3}$$

where α_{gel}, α_{spacer} are the conductivities (*A/V*) in the gel and the spacer respectively, and Δw is the spacer width which is assumed to be small compared to its length. The first term of the boundary conditions are derived from the orthogonal field component and the second term originates from the leakage current in the imperfect gluing of the spacer. If we assume the migration velocity of the SDS-protein complexes ν (m/s) is a linear function of the electric field then:

$$\nu = -\beta\nabla\Psi \tag{4}$$

where β is the mobility coefficient (m^2/Vs). From Eqs. (2) and (4) and also noticing that time can be expressed in terms of settled distance leads us to a set of differential equations that specify the function *m* mapping each distorted pixel location to a corrected pixel location:

$$\frac{\partial}{\partial\gamma}m(x,y) = -\frac{h}{V_0}\nabla\Psi(m(x,y)) \tag{5}$$

for $y = 0$ at the cathode to $y = c$ at the ideal gel front. *h* is the gel height. To use the current leakage model one has to correct for a global scaling, rotation and translation and then estimate the parameters of the model. This normally involves fitting a straight line to the cathode and a low degree polynomial to the gel front. This fixes the scaling and rotation and vertical translation, but not the horizontal translation ξ. So there are now four unknowns:

$$\Theta = \{c, h\gamma_L, h\gamma_R, \xi\} \tag{6}$$

This space of model gel front curves is searched for the optimal least squares fit of the actual gel front. Once found a 16 × 16 grid is defined from Eq. (5) and the gel warped by it (Fig. 2). One drawback in that unfortunately some labs run their gels so that the gel front diffuses off the image.

If an ending condition for the idealized gel can be found this process could be automated. For instance, notice that in an ideal gel if each spot is modelled as an ellipse its half-axes will lie along the coordinate axes. In a gel with current leakage each ellipsis's half-axes will lie tangent and normal to the current leakage mapping. A set of ellipses would lead to a unique solution to the model's parameters in this way.

2.1.2 Intensity correction

Through their acquisition and susceptibility to dust, most images, including 2-D gels, need to be smoothed to suppress the statistical Gaussian noise inherent in them. The most common smoothing technique is a local ($n \times n$ window) Gaussian, diffusion or polynomial convolution filter [34], or a local median filter. Histogram equalization and contrast enhancement redefines the intensity values in the image to obtain higher contrasts between spots and background [39, 34].

Background subtraction is applied to eliminate meaningless changes in the gel background intensity level. A simple approach is to obtain the lightest and darkest point in the background and replace the whole background with the average intensity. Tyson and Haralick [47] find the local minima in the image, representing background depressions, and interpolate the background between these minima. Melanie II [34] subtracts the minimum

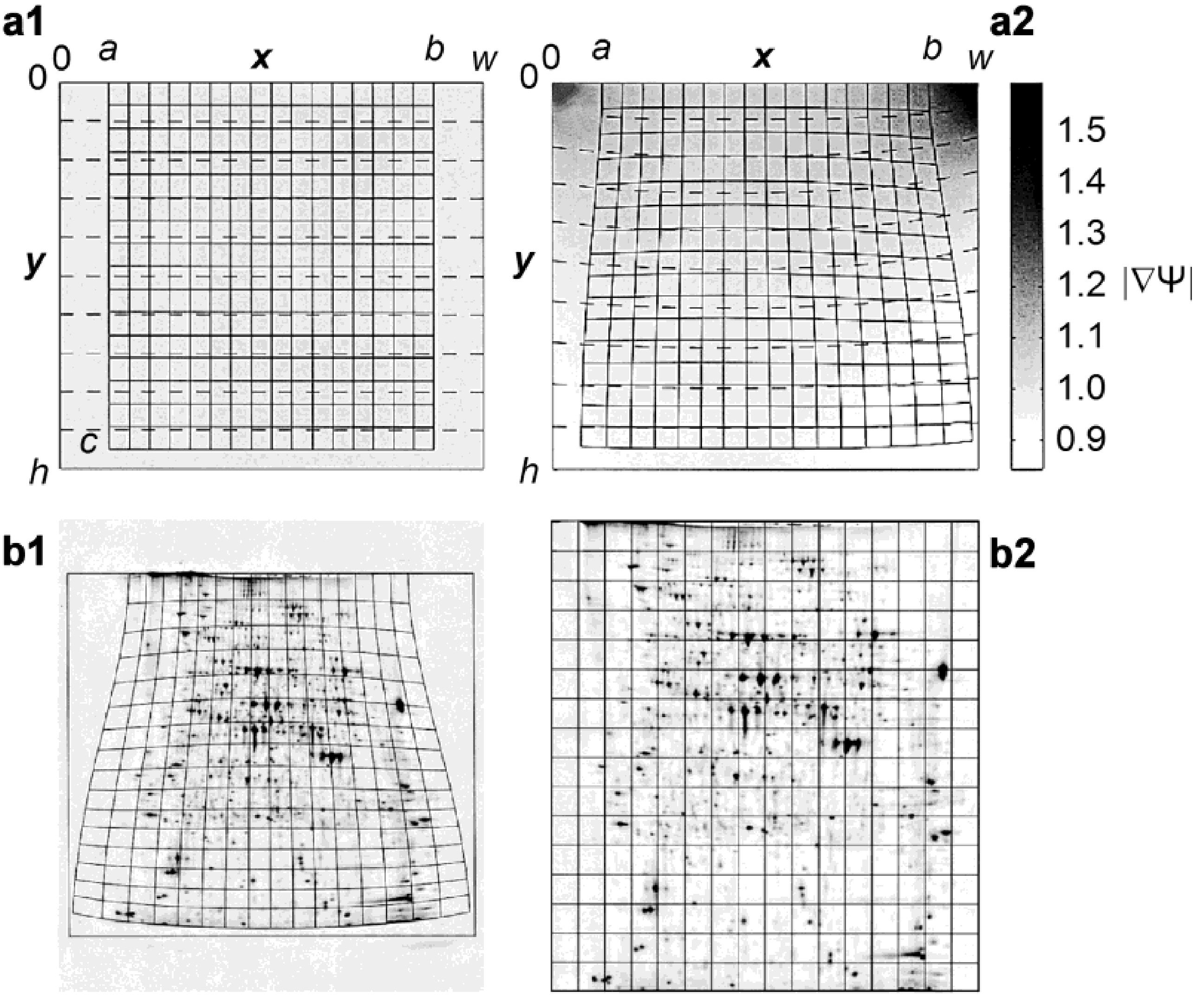

Figure 2. Current leakage correction of Gustafsson *et al.* [46]. (a1) Current leakage mapping for an ideal gel; and (a2) for a gel with current leakage across the sides. Equipotential lines are drawn as dashed curves and the electric field strength at each point is given by the greyscale intensity. The scale, given in the bar to the right, is relative to the undisturbed field strength in the ideal gel. The dimensionless conductivity parameters $h\gamma_L$ and $h\gamma_R$ are 0.2 and 0.4 respectively. (b1) shows an example gel with the detected current leakage mapping superimposed; (b2) is the transformed gel warped by the mapping. In this case $h\gamma_L$ = 0.65 and $h\gamma_R$ = 0.7. Reproduced from [46], with permission.

intensity from all pixel values and then fits a third degree polynomial to the background image (with spots removed). Another technique comes from 3-D mathematical morphology, where the operations of opening and closing a greyscale image by a structuring element is represented by sliding the structuring element under and over the topological image respectively (intensity is regarded as height) [48]. Background variability in background subtraction is estimated by opening the image with a spherical structuring element 'the rolling ball', which is larger than the largest spot but with more curvature than the background. As the ball rolls under the image it is prevented from entering the narrow crevices of the spots. Each pixel height in the resultant background image is the highest point the ball can reach.

Opening is defined as erosion followed by dilation, and closing defined as dilation followed by erosion. The dilation of an image is the topography produced by the maxima of the structuring element touching every point of the image, and erosion the minima. By using a horizontal or vertical cylindrical structuring element horizontal or vertical streaks may also be removed [49]. The cylinder rolls into streaks but cannot roll into spots, therefore the output is an image of streaks that can be subtracted from the original image.

2.2 Spot detection

Spot detection concerns the individual resolution of each spot, thereby outputting a list of spot centers, intensities and geometric properties. Since spots vary greatly in size and intensity they are difficult to distinguish from artefacts such as noise and streaking and can overlap other spots. Furthermore, heavily saturated clusters may have no observable boundaries between spots, so the task is not trivial. The current spot detection pipeline in most cases is: (a) Detect the centers of as many spots as possible; (b) Segment the gel into regions each containing one of these spots; (c) Model each region by a prior parametric spot model to both (1) extract a characteristic vector for each spot for data analysis, and (2) detect and separate comigrated spots.

Early methods used a wide variety of techniques. Laplacian of Gaussian (LoG) techniques [21, 31, 33] detect the spot centers as the crossings of the second derivative image and find an initial 'core' segmentation where it is negative. LoG's sensitivity to noise requires the image to be Gaussian prefiltered. The core areas are then propagated to neighboring pixels using heuristics based on the intensity values and second derivatives [50], or by fitting a 2-D polynomial to the core [31].

PiKA^2 uses the ring operator [36] instead, which peaks in the center of ellipses after an initial image thresholding into spot and background regions. The line and chain analysis algorithms of Garrels [51] find the peaks in each vertical scanline. The peaks on neighboring scanlines are then chained to both determine the spot centers and segment the image.

Prehm *et al.* [52] directly segment the gel by assuming the spots belong to areas with convex intensity curvature, calculated by convolving the image with an $n \times n$ spot kernel. The areas between overlapping spots do not show as concave, so they will not be separated. Instead, the difference in convexity along four directions through the test pixel is sampled (by calculating the convexities m pixels away from the test pixel) and a spot detected if for at least one direction it is above a threshold. The ideal sizes of n and m are dependent on the resolution of the gel image. Conradsen and Pedersen [53] use morphological operations equivalently. A series of filters with increasing kernels (3×3, 5×5, 7×7, 9×9) are used to cope with various spot sizes. Median noise removal is interleaved with local-maximum erosion (to separate overlapping spots) and second derivative edge detection, which outputs a binary segmentation image. Working backwards, the binary images are processed with interleaved summing and binary erosion operations. Summing fills in large spots that the edge detection outputs as rings, and the erosion makes sure the spots do not re-overlap. The erosion will remove small spots but the information still remains in the previous binary image.

Another option is the h-basin transformation which Horgen and Glasbey [54] applied to 2-D gels. Regional maxima not exceeding h in size are found by first subtracting h from all pixels in the image. Iterative geodesic dilations are then performed on the resultant image until stability, in other words the heights of the maxima propagate outwards until they collide. When this image is subtracted from the original images only the maxima domes remain.

Currently, the most popular technique for spot segmentation is the watershed transform (WST) due to its robustness to noise. Watersheds [55] are a terminology from geoscience. A watershed is the boundary of a region (catchment basin) in a landscape where all water drains to a common point. We treat the gel as a topographic relief, where the protein spots are depressions. The watershed transform assigns labels to the pixels in a gel such that different catchment basins are uniquely labelled and a special label W is assigned to pixels of the watershed.

The most popular algorithmic approach by Vincent and Soille [56] follows the immersion principle. Imagine piercing all local minima and slowly submerging the landscape in water. Where catchment basins merge dams are built and watershed points are labelled (Fig. 3). The algorithm first sorts the image pixels by ascending height, and then follows an iterative flooding step. At each height h starting at the lowest point in the image h_{min} all pixels with height h are marked. If a marked pixel has a labelled neighbor the labelling is recursively propagated outwards to all connected marked pixels. If a marked pixel has two differently labelled neighbors we have found a watershed. Any unassigned marked pixels left over are assigned as new catchment basins.

The watershed transform lends itself well to gel segmentation [57] as the spots are characterized by monotonically increasing and then decreasing shape. One major disadvantage is the tendency for over segmentation due to noise creating false minima. There are two main solutions: (a) Remove unwanted catchment basins before segmentation (marker controlled watersheds). (b) Remove unwanted watersheds after segmentation (region merging). In the marker controlled WST only selected

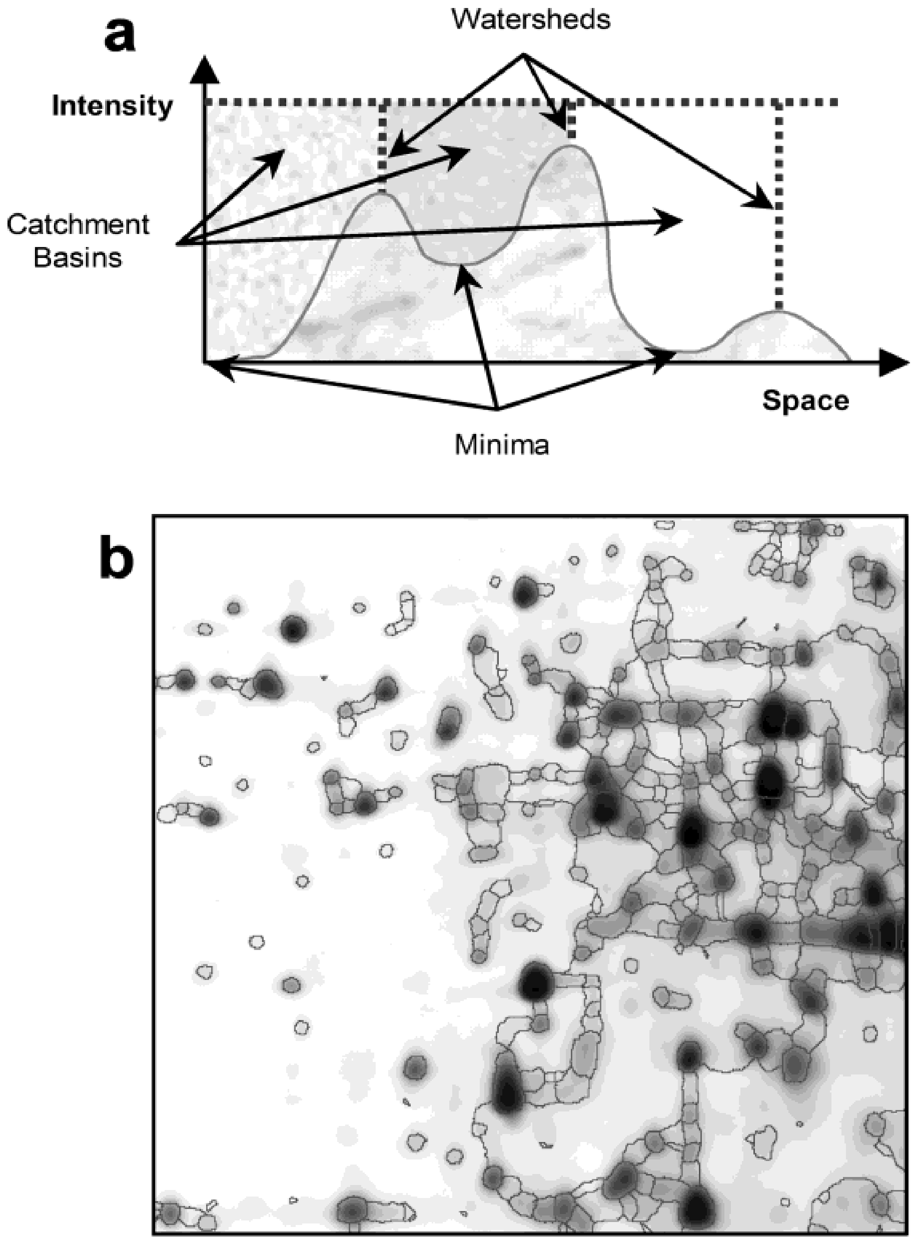

Figure 3. The watershed transform. (a): Cross-section of a region processed by the watershed transform (WST), treated as the immersion principle of Vincent and Soille. The minima are pierced and the region slowly submerged in water. Where catchment basins meet watershed 'dams' are constructed. (b): Gel segmented by the WST on the gradient image. Note: watersheds are built where the first derivatives are locally greatest.

minima are pierced. Image homotopy modification [58] is then used to suppress all other minima by filling in their catchment basins.

Pleissner *et al.* [59] perform region-merged WST on the gradient image (the first derivative). This assumes the optimal spot contours will appear as ridges. However, no longer will the WST differentiate between background regions and spots. To do this, two thresholds are firstly used: the mean intensity of a spot region will be substantially higher than at least one of its (background) neighbors; if the spot region is land-locked by other spot regions, a simple intensity threshold is used instead. Secondly, since spots (and partial spots) have convex curvature, regions are merged that satisfy the condition $C(r) > 0$:

$$\mathrm{C}(r) = \sum f''(p) \qquad (7)$$

where the sum is taken over all pixels p in the region and f'' is the second derivative image.

Recently Baker *et al.* [38] use all the information in a stack of preregistered (see Section 3.1.3) gels of the same experiment to allow the noise in the image to be modelled so that even faint spots can be identified. Their segmentation algorithm uses a Markov Random Field (MRF) approach based on the Laplacian.

MRF's [60] is a branch of probability theory for analyzing the spatial dependencies of physical phenomena. A random field $F = \{F_1, \ldots, F_n\})$ on S is a family of random variables, where $F \in S$ and each F_i takes a value f_i from the set of labels L. For discrete images each random variable is a pixel location. A neighbourhood system N is incorporated so that each pixel is connected to its four nearest neighbors, thus the field becomes a graph (or lattice). A MRF with respect to N must have these properties:

$$P(f) > 0 \, \forall f \in F$$
$$P(f, |f_{\mathrm{s}-(\mathrm{i})}) = P(f_{\mathrm{i}}|\{f_{\mathrm{j}}|j \in N_{\mathrm{i}}\}) \tag{8}$$

The second equation, Markovanity, states each labelling must only be dependent on the labelling of the pixel's neighbors. MRF's are ideally specified by its joint probabilities *P(f)*, whose calculation is tractable since MRF's have been found to be equivalent to Gibbs Random Fields (GRF). GRF's are characterized by the global condition of Gibbs distribution (rather than the local condition characterizing MRF's):

$$P(f) = \frac{e^{-\frac{i}{T}\mathrm{U(f)}}}{\sum\limits_{\mathrm{f} \in \mathrm{F}} e^{-\frac{i}{T}\mathrm{U(f)}}} \tag{9}$$

where T is a constant temperature, usually 1, and *U*(*f*) the energy function. Note the Gaussian distribution is a special member of the Gibbs distribution family.

Baker *et al.* [38] first compute the binary core images C_m, where 1 represents a positive Laplacian in either dimension. Spots not appearing in multiple images are penalized by minimizing the follow energy function with the Iterated Condition Mode (ICM) relaxation scheme [61]:

$$\begin{aligned} U(S,x,y) &= (1 - S(x,y)) \sum_{\mathrm{k}=1}^{\mathrm{m}} C_1(x,y) \\ &+ b_{\mathrm{M}} S(x,y) \sum_{\mathrm{m}=1}^{\mathrm{M}} [4 - S(x-1,y) - S(x+1,y) \\ &- S(x,y-1) - S(x,y+1)] \end{aligned} \tag{10}$$

where S is the single binary segmentation image. In the second stage another penalty function was minimized. This MRF modelled Wu *et al.*'s [50] spot growing conditional – for a pixel labelled 0:

$$\exists (x',y') \in N(x,y) \wedge S(x',y') = 1$$
$$\wedge I(x',y') < I(x,y) \wedge ||\nabla I(x',y')|| > ||\nabla I(x,y)|| \tag{11}$$

An issue with the above techniques is that spots can overlap in a way that only one segment is found, so greater prior information of spot shape is required to differentiate between them. A parametric spot model is a functional description of an idealized spot with parameters θ. Assuming all spots have common characteristics that can be modelled, simplified spot matching and quantification can result when card(θ) ≪ card(*I*). Fitting each spot to the model concerns taking a segment Ω of *I* containing one spot and using an optimizer to minimize function *f*, the squared residuals of θ:

$$\min_{\Theta} \left(f_{\Theta}(w) = \sum_{(\mathrm{x,y}) \in \Omega_{\theta}} [G(x,y) - I(x,y)]^2 \right) \tag{12}$$

Most current techniques [21, 22, 26, 28] involve fitting with 2-D Gaussians:

$$G(x,y) = e^{\mathrm{ax^2+bxy+cy^2+dx+ey+f}} + h[2b - ac < 0] \tag{13}$$

where θ = {*a, b, c, d, e, f*}. More intuitively:

$$G(x,y) = h + A\exp\left(-\frac{(x-x_0)^2}{2\sigma_{\mathrm{x}}^2}\right)\exp\left(-\frac{(y-y_0)^2}{2\sigma_{\mathrm{y}}^2}\right) \tag{14}$$

where θ = {*A*, *B*, x_0, y_0, σ_x, σ_y}. (x_0, y_0) is the Gaussian's center, *A* its amplitude, *h* the background intensity and (σ_x, σ_y) its diffusion coefficients along the principle axes. Melanie II [34] uses the Polak-Ribiere variant of the conjugate gradient method to optimize the parameters, which requires the partial derivatives of *f* with respect to θ. The convergence of the iteration depends greatly on the starting estimate. A good starting approximation for Eq. (13) assumes *h* is constant and always less than *I* (*x*, *y*). Equation (13) then reduces to the following set of linear equations for $(x, y) \in \Omega$

$$ax^2 + bxy + cy^2 + dx + ey + f = \log(I(x,y) - h) \tag{15}$$

Bettens *et al.* [62] noticed that when the local concentration of protein is high, saturation effects occur and the spot no longer can be accurately modelled by a Gaussian. Instead, they model the spot by a simplified diffusion process. The fundamental differential equation for diffusion in an isotropic 2-D medium is given by:

$$\frac{\partial C}{\partial t} = D\left(\frac{\partial^2 C}{\partial^2 x} + \frac{\partial^2 C}{\partial^2 y}\right) \tag{16}$$

where *C*(*x,y*) is the concentration of the substance at point (*x,y*), and *D* the diffusion constant. Rewriting Eq. (16) in polar coordinates and solving the differential equation leads us to the equation specifying *C* for the radius of diffusion *r* at time *t*, with *M* the total amount of the diffusing substance:

$$Cr = \frac{M}{2\sqrt{\pi D t}} \exp\left(\frac{-r^2}{4Dt}\right) \tag{17}$$

In 2-D PAGE the complex initial distribution of protein is estimated as being uniformly distributed within a circle of radius *a*. Also separate diffusion constants D_x and D_y are incorporated for each dimension as the diffusion is anisotrophic. The final diffusion model is found by removing symmetric elements and adding two extra position coordinates:

$$C(x,y) = B + \frac{1}{2}C_0\left[\operatorname{erf}\left(\frac{a'+r'}{2}\right) + \operatorname{erf}\left(\frac{a'-r'}{2}\right)\right] + \frac{C_0}{r'}\sqrt{\frac{1}{\Pi}}\left[\exp\left(-\frac{(a'+r')^2}{4}\right) - \exp\left(-\frac{(a'-r')^2}{4}\right)\right] \quad (18)$$

where $r' = \sqrt{\frac{(x-x_0)^2}{D'_x} + \frac{(y-y_0)^2}{D'_y}}$

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt \qquad a' = \sqrt{\frac{D}{t}}a$$

$$D'_x = D_x t \qquad D'_y = D_y t$$

C_0 is the initial concentration in the circle and erf is an error function. $\theta = \{B, C_0, x_0, y_0, a', Dx', Dy'\}$. Notice that for small *a* Eq. (18) reduces to the Gaussian model Eq. (14) (see Fig. 4 for a visual comparison between the Gaussian and diffusion models).

Rogers *et al.* [45] recently proposed a parametric model based on the statistics of spot shape observed from an annotated training set, convolved with a Gaussian kernel. The model is derived backwards from observed spots rather than forward from spot formation idealizations. It is

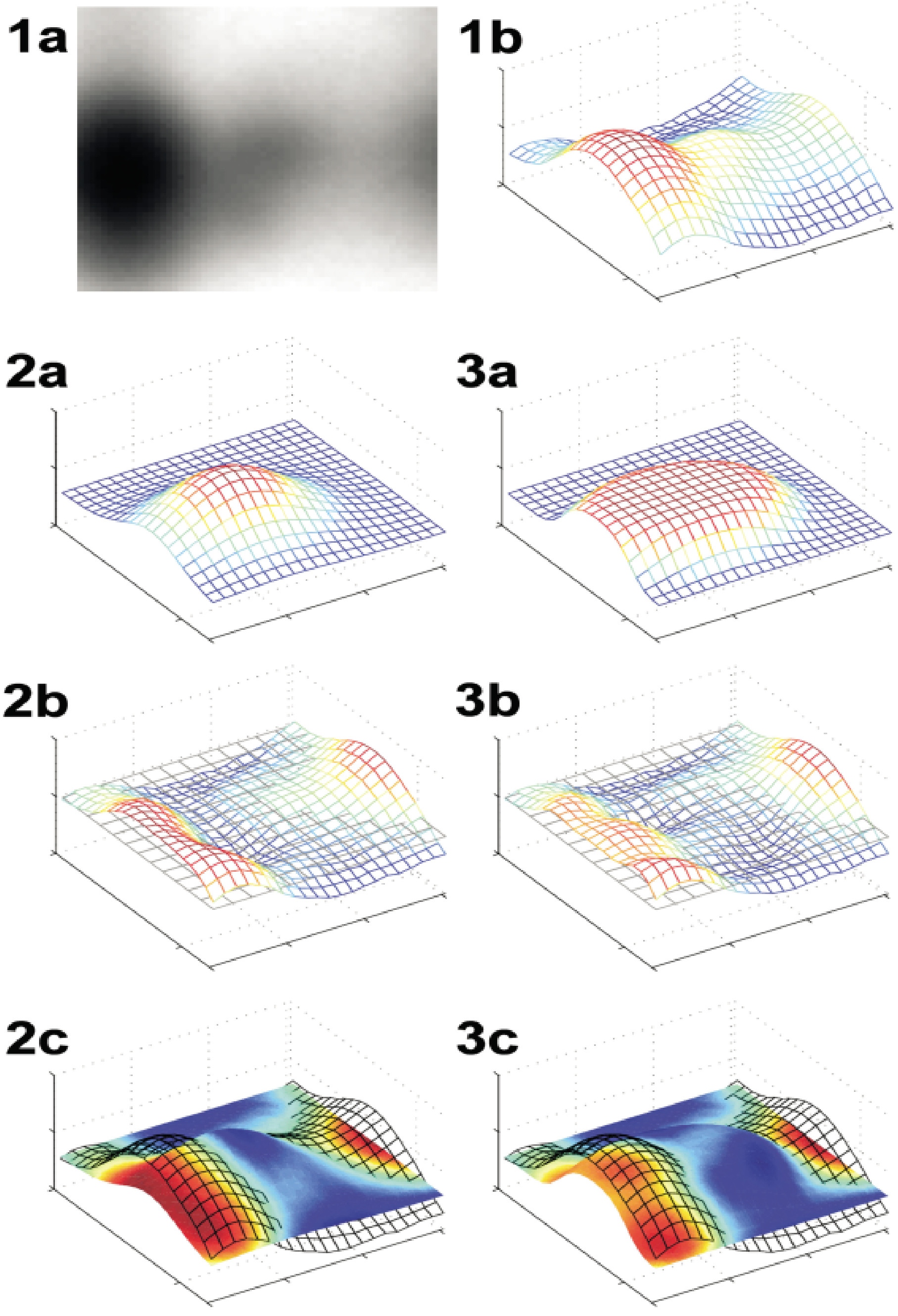

Figure 4. Parametric spot fitting of the shoulder spot seen in the middle of image (1a). (2a) is the image's wireframe representation. Gaussian and diffusion modelling are illustrated in series b and c respectively. (1b–1c) show the fitted model; (2b–2c) show the residuals after the fitted model is subtracted, (3b–3c) show the image's wireframe representation overlayed with the fitted model as a solid surface and coloured depending on the residual. Reproduced from [78], with permission.

able to detect unusually shaped spots but is still specific enough to highlight complex regions. The shape modelling is performed using a point distribution model (PDM). Forty evenly spaced points are placed on the perimeter of each training spot. The training data is processed by principle component analysis (PCA) [63] to find the principle modes of variation.

PCA analyses the interrelationships among a large number of variables and explains them in terms of their common underlying factors *i.e.* orthogonal dimensions. The first principle component is the linear combination of the variables with maximal variance *i.e.* contains the most information about the data. The second principle component has maximal variance subject to be being orthogonal to the first, and so on. Principle components are computed as the eigenvectors of the covariance matrix. Standardized PCA uses the correlation matrix equivalently. The eigenvectors are found using singular value decomposition (SVD) and then ordered in eigenvalue order. PCA is often used as a data reduction technique as the majority of the data is summarized in less dimensions.

With PCA, any example *x* from the training data can be approximated by its variance from the mean:

$$x = \bar{x} + P\theta \tag{19}$$

where *P* is the matrix of the first few eigenvectors covering 95% all variation and θ the vector of model parameters. Fitting a spot is a process of optimizing θ for a new *x.*

One unanswered question is how to make sure each segment contains a single spot for fitting. In [62] the watershed transform of Vincent and Soille [56] is performed prior to fitting each segment. If a single spot fit *C* is subtracted from I_w, spikes in the output will indicate the presence of tertiary spots. A spot model equivalent to the mixture of two or more spot models would then be used on I_w instead, the improvement in fitting verified by a χ^2 test.

Lemkin *et al.* [64] realised that due to wide saturation effects in complex regions, intensity information will not help for segmentation. The merged spots are cut at opposing saddlepoints (concavities in the boundary). The algorithm first finds all robust concavities and then tries to match complementary ones. Efrat *et al.* [65] propose that spot detection in these areas is an ellipse covering problem. The space of possible ellipses is reduced by adhering to the following four restrictions: (a) Shape: The ratio of the half-axes are in the interval $[1/\alpha, \alpha]$, with threshold $\alpha \geq 1$. (b) Fitting: The ellipse is fully contained within the complex region. (c) Intersection: The boundaries of each pair of ellipses only intersect at a maximum of two points. (d) Coverage: At least β% of the complex region is covered by ellipses.

In the brute force approach the space of ellipses are also discretized. For each pixel in the complex region a group of ellipses are generated whose centers are the pixel center. In the group there is one ellipse for each *x* half-axis length that both fits in the complex region and is a multiple of the pixel side length. The *y* half-axes are then chosen as the maximum values that satisfy the fitting constraints. A greedy approach is then used to find the minimum set of ellipses adhering to the intersection and covering constraints. Because the discretization can miss optimal ellipse coverages, and generates many unnecessary ellipses, Efrat also describe a linear programming approach in their paper. A triplet of pixels is chosen randomly from the complex region until an ellipse containing them can be fitted. The algorithm continues in a Metropolis methodology, randomly adding neighbors into, and deleting from, the ellipse containment test. After a set number of rounds the generated ellipse with half-axes ratio closest to 1 is added as a candidate for the covering.

2.3 Spot analysis (quantification)

After spot detection, characteristic information about each spot is extracted both to quantify the protein expression and to aid in the spot matching process. Examples from [24, 34] include:

Spot Area:

$$\text{AREA} = \text{number of pixels} \times \text{pixel area} \tag{20}$$

Spot Optical Density:

$$\text{OD} = \max_{(x,\,y)\in\,\text{spot}} I(x,y) \tag{21}$$

Integrated Optical Density:

$$\text{VOL} = \sum_{(x,\,y)\in\,\text{spot}} I(x,y) \tag{22}$$

Further statistics can be derived from the parameters of the parametric spot model [28].

Integrated Spot Intensity:

$$\text{ISI} = \pi A \sigma_x \sigma_y \tag{23}$$

Integrated Gaussian Density:

$$\text{VOL} = \sum_{(x,\,y)\in\,\text{spot}} G(x,y) \tag{24}$$

%VOL and %OD (normalized by the total over the gel) are typically used to quantify protein expression, though %VOL has been shown to be more accurate [66]. Neither account for background stain levels and both have a limited range of linearity. In a recent investigation into quantifying silver stained expression a new statistic, the scaled volume (*SV*), was devised [66]:

$$\mathrm{SV} = \frac{\mathrm{VOL}_{\text{spots of interest}}}{\left(\frac{\mathrm{VOL}_{\mathrm{gel}} - \mathrm{VOL}_{\text{spots not of interest}}}{\mathrm{AREA}_{\mathrm{gel}} - \mathrm{AREA}_{\text{spots not of interest}}}\right)} \tag{25}$$

The *SV* represents volume of a spot normalized with respect to background. In an experiment containing three protein samples run at 10^2, 10^3 and 10^6 ng *per* gel, the *SV* was found to increase exponentially with protein amount, and was invariant of developing time.

3 Differential image analysis

3.1 Image registration

In many disciplines, comparison needs to be made for images from different sensors and modalities, at different times or conditions. Image registration represents a class of techniques to align and normalize two images 'optimally'. In this sense 'optimally' depends on what needs to be matched. Usually geometric and intensity distortion attributed to the acquisition of the images (*e.g.* sensors in different locations) or experimental uncertainty (*e.g.* protein diffusion in 2-DE) are undesirable and therefore should be eliminated by image registration. On the other hand, intrinsic difference (*e.g.* different modalities, brain legions, changes in protein expression) must be kept.

The problem is a quadruple (I_s, I_r, M, *sim*) where the source image I_s must be brought into alignment with the reference image I_r, constrained by the space of allowable transformations (mappings) on I_s, M, and guided by the similarity measure function *sim* – which is at a maximum when the alignment is optimal:

$$\arg\max_{m} sim(I_s \circ m, I_r) \text{ where } m \in M \tag{26}$$

The maximum is found with an optimization strategy on the parameters θ of the mapping function m, which transforms each pixel location in the sample image space to a location in the reference image space.

Reviews of general image registration, medical image registration and similarity measures can be found in [67–69] respectively. The nature of M is very important as there is a trade-off between finding the transformation with the greatest similarity and insisting on a smooth transformation. The prior often maximizes the efficiency (proportion of found matches to total matches), whilst the later aims to maximize the accuracy (proportion of found matches to true matches).

In 2-DE registration has operated on the spot features or directly on the pixel values, either automatically or with manual intervention (landmarking). Spot matching concerns the automatic pairing of spot lists in the reference and sample gels using the data from the spot detection phase. Landmarked 'seed' pairs are sometimes required. We define a match vector as the directional vector linking the centers of a spot pairing, then the similarity measure is greatest usually when the sum of match vector magnitudes are smallest. To cope with deformations, algorithms usually just rely on a robust search strategy, but in recent times a direct image transformation stage has been incorporated for more reasonable modelling of the distortions. Either way the output is the optimum set of match vectors.

Hybrid and direct methods have appeared in recent times, in which the similarity measure contains some weighting for pixel level correlation. The end result is an optimal transformation mapping which the sample gel is warped by. Much less burden is then placed on the spot matching algorithm in the next phase.

Another issue in spatial registration is the emerging use of narrow and very narrow-range IPG's for improved protein resolvability in complex proteomes [9]. It is necessary for software packages to piece together gels of the same sample but different (overlapping) p*I* ranges. This is essentially an extension of image registration but with a much greater global translation component than is usually allowed. For example, PDQuest requires landmarks be supplied to perform this.

3.1.1 Image warping

It is worth mentioning image warping [70] before introducing registration techniques as a good proportion explicitly transform the pixel information by m at some point. Also, manual image warping with landmarks is the image registration technique traditionally used as a precursor to robust spot matching. Therefore warping is explained in this context.

The number of parameters in the transformation model determines the number of landmarks required for a unique warp using linear regression, or alternatively the similarity measure can be optimized over the transformation parameter space. Many systems [39, 34, 71, 72] apply the n-order polynomial transform. They use the least squares method to solve the following two systems of linear equations:

$$q_x = \sum_{i=0}^{n} \sum_{j=0}^{n-1} \alpha_{i,j} p_x^i p_y^j$$
$$q_x = \sum_{i=0}^{n} \sum_{j=0}^{n-1} \beta_{i,j} p_x^i p_y^j \qquad (27)$$

where Δpq are the match vectors and $\alpha_{i,j}$ and $\beta_{i,j}$ are the polynomial coefficients to be determined. These parameters $\Theta = \{\alpha_{1,1}, \beta_{1,1} \ldots \alpha_{n,n}, \beta_{n,n}\}$ determine the mapping:

$$m_\Theta(x,y) = \left(\sum_{i=0}^{n} \sum_{j=0}^{n-i} \alpha_{i,j} x^i y^j, \sum_{i=0}^{n} \sum_{j=0}^{n-i} \beta_{i,j} x^i y^j \right) \qquad (28)$$

$(n + 2)(n + 1)/2$ match vectors are required to find a unique solution.

Horgan *et al.* [73] compared affine and thin plate spline (TPS) transformations whilst Pedersen [72] compared bilinear mapping and TPS. A bilinear mapping maps a square in the reference image to a quadrilateral in the sample image where the parameters (control points) are the four corners of the quadrilateral:

$$m_\Theta(x,y) = (1-u)(1-v)c_{0,0} + u(1-v)c_{1,0} + $$
$$+ v(1-u)c_{0,1} + uvc_{1,1}$$
$$\Theta = \left\{ c^x_{0,0}, c^y_{0,0}, c^x_{1,0}, c^y_{1,0}, c^x_{0,1}, c^y_{0,1}, c^x_{1,1}, c^y_{1,1}, \right\} \qquad (29)$$

u is the ratio $c_{0,0}$:$c_{0,1}$ and $c_{1,0}$:$c_{1,1}$. v is the ratio $c_{0,0}$:$c_{1,0}$ and $c_{0,1}$:$c_{1,1}$. Four non-degenerate match vectors are required to uniquely identify the eight parameters. A TPS transformation is an arbitrary mapping constrained by penalising the similarity measure with a smoothness constraint:

$$\mathrm{tps}(I_r, I_s) = -\sum_{i=1}^{n} ||p_i - m(q_i)||^2 -$$
$$- \lambda \iint_{I_x\, I_y} \left(\left(\frac{\partial m^2}{\partial^2 x}\right)^2 + \left(\frac{\partial m^2}{\partial x \partial y}\right)^2 + \left(\frac{\partial m^2}{\partial y^2}\right)^2 \right) dxdy \qquad (30)$$

where the first term is the summed magnitude of match vectors Δpq transformed by m and the second term imposes a limit on the second partial derivatives of *m i.e.* a constraint on the bending energy of the spline. Both papers found the TPS transformation lead to a superior match. Pedersen noted that the TPS could capture all the information from any number of landmarks, whereas a bilinear mapping could not.

For this reason Salmi *et al.* [74] proposed a multiresolution piecewise bilinear mapping approach. A piecewise bilinear mapping maps one grid of convex quadrilaterals in the reference image to another in the sample image. Therefore each non-boundary control point is shared between four quadrilaterals. In a multiresolution approach the corners of a single quadrilateral covering the image is optimized, which is then recursively subdivided and optimized to correct for finer and more local distortions. The bounding box parallelograms are first constructed for both images and then at each stage the control points are optimized using a simple descent method on the summed magnitude of the landmark match vectors.

3.1.2 Feature based methods (spot matching)

Feature based registration relies on the similarity of geometrical features extracted from the reference and sample images. In 2-DE this is a list of spot centers annotated by scalars such as *IOD* Eq. (22). Researchers have adapted techniques from computational geometry to compensate for geometric distortions in the spot matching phase, either automatically or with a prior landmark warping stage. Akutsu *et al.* [75] proved that spot matching under non-uniform distortion is NP-hard in two or more dimensions. Explained, P is the set of decision problems solvable in polynomial time. NP problems can only for certain be verified in polynomial time, their computation is non-deterministic polynomial time. NP-complete problems are the hardest problems in NP, and are all (so far) exponential time *e.g.* is there an *n* spot matching with similarity error $<\varepsilon$? Finally, NP-hard problems are optimization versions of NP-complete problems *e.g.* what is the maximum cardinality spot matching with error $<\varepsilon$? The only way to improve the performance of a NP-hard problem is to use an approximation, probabilistic or heuristic based approach.

The well researched technique of point pattern matching (PPM) [76] is the standard approach used today. Its objective is to find all occurrences of a finite point pattern *P* in a target point set *T*. We must find a transformation mapping each point in the pattern set on, or as close as possible, to a point in the target set. The approximation case requires a match between *P* and $Q \subseteq T$ to have a similarity under an error threshold. The most common similarity measure is the Hausdorff distance [77]:

$$H(P,Q) = \min\,(\widetilde{H}(P,Q), \widetilde{H}(P,Q))$$
$$\text{where } \widetilde{H}(P,Q) = \max_{p \in P} \min_{q \in Q} d(p,q) \qquad (31)$$

where *d* is a distance metric, usually the Euclidian distance. In words, the Hausdorff distance assigns to each point of one set the distance to its closest point in the other and takes the maximum over all these values.

In 2-DE the pattern P_s is the set of spot centers in I_s, and the target set P_r is the set of spot centers in I_r. There should be one only occurrence of the pattern in the target set – the true spot matching. Due to difficulties in resolving all spots in the spot detection phase, and differential protein expression across the gels, finding a bijection between the sets is unlikely. It is desirable that these methods should [78]: (a) Exactly and robustly match protein pairs; (b) Allow for non-linear distortions; (c) Robustly handle outliers in both sets; (d) Be able to handle point sets of stochastic nature; (e) Robustly match dense point sets. Algorithms have traditionally met the first three targets, and advances in tackling the rest have become available recently.

A wide variety of fully PPM methods have been developed such as: iterative closet point (ICP), the alignment method, bipartite graph matching (geometric hashing), expectation maximization (EM), dynamic programming and robust pattern matching (RPM). Of these, EM has not been used for 2-D gel spot matching, but is described in Section 3.1.4 for intensity registration. Only RPM incorporates explicit image warping directly into the point matching algorithm.

The alignment method is based on the observation that any rigid transformation (except reflection) is determined by the mapping of a single line segment (arc) between two points in a point set. A set of candidate transforms can be generated by mapping any arc in the pattern on to all arcs (the fully connected graph) in the target. The pattern is transformed by each candidate and the similarity calculated. The one with the highest similarity is the optimum match. Unfortunately every arc in the pattern must be mapped on to every arc in the target set as the best partial match is required, a bijection between $Q_s \subseteq P_s$ and $Q_r \subseteq P_r$. For k = card(P_s) and n = card(P_r) the worst case upper bound is (order) O(k^2n^2) arc pairs and O(k log n) for computing the pattern similarity of each pair.

The use of prior knowledge can greatly reduce the alignment method's search space. We can significantly restrict the allowed rotation, scaling and spot intensity difference of the matched arcs. Another optimization is to remove a subset of the arcs as redundant, a trade-off between efficiency and robustness. One way is the use of proximity graphs (beta-skeletons) [79], where points are only connected if no other points lie within a certain region. By using beta-skeletons the arc count rises linearly with the point count, so that the arc pair tests becomes O(k^2n).

Garrels [51] used the beta-skeleton Gabriel Graph (GG) as a basis for spot matching. In a GG points are connected when no other point lies within the circle containing the arc as its diameter. PiKA^2 [36, 37] uses a Relative Neighbourhood Graph (RNG) for the reference pattern and a Delaunay Net (DN) for the sample pattern. In an RNG, a subset of a GG, points are connected if no spot lies in the intersection of two circles centered on the two points whose radii are the arc. A triangulation is a maximum set of non-intersecting line segments (tessellation) that make up only triangles. A Delaunay Net, a superset of a GG, ensures that only the three corner points of each triangle lie within its circumcircle (Fig. 5). PiKA^2's intention is that the RNG is relatively transformation insensitive compared to the DN. Automatic matching proceeds with the alignment method.

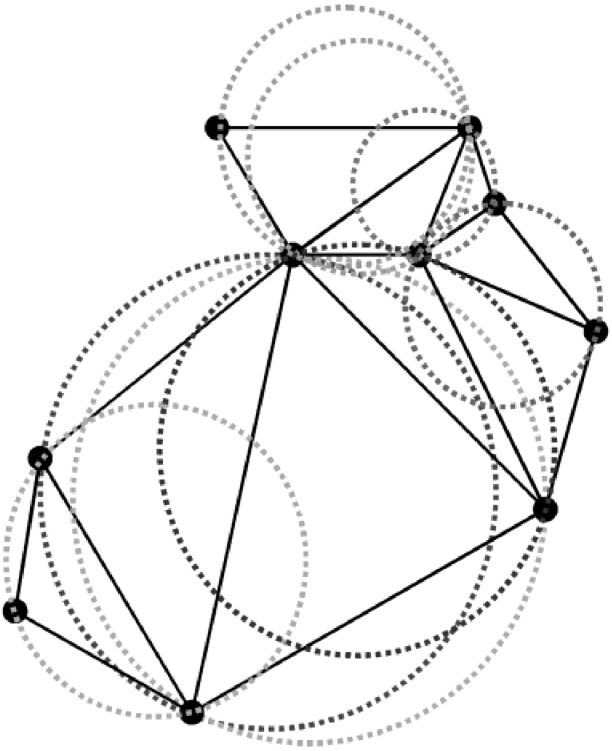

Figure 5. Delaunay triangulation of a point set. No other point other than the three vertices must be contained within the circumcircle (dotted) of each triangle.

Early methods [21, 28] are based on iterative closest point (ICP) methods – successive point matching propagated outwards from landmarks. Points are matched either by checking that local clusters match reasonably [31] or assigning a confidence to the match dependent on its distance from a user set landmark [24]. Many matches are found and the method selects the most consistent. Any unmatched spots are transformed relative to the set of matched spots before the algorithm repeats. LIPS [29] constructs two GG's and iteratively selects a pair of spots neighboring the previous match. The new match vector is transformed by the negation of the previous match vector, and if its magnitude is small, the spots are a match. Melanie II [34] finds each ideal match vector as the maximum probability of randomly matching m or more spots in the surrounding spot clusters with N tries and m the number of spots in one of the clusters:

$$P(m) = \sum_{\mathrm{h=m}}^{\mathrm{N}} \left[\binom{N}{h} \prod_{\mathrm{i=1}}^{\mathrm{h}} \rho_{\mathrm{i}} \prod_{\mathrm{i=h+1}}^{\mathrm{N}} N(1 - \rho_{\mathrm{i}}) \right] \tag{32}$$

$$\text{where} \quad \rho_m = \frac{A_s - A_{m-1}}{A_c - A_{m-1}}$$

This is approximated by N Bernoulli trials of m random dart throws. For consistency the clusters are checked against a limited range of admissible rigid transformations.

Geometric hashing is a technique developed in computer vision for efficiently matching geometric models against a database of such models [80]. A quantized 2-D hash table covers a reference coordinate system. Each model pattern $Q_r \subseteq P_r$ is preprocessed by computing the following: For every minimal geometric feature in Q_r able to unambiguously define m (for rotation, translation and scaling this is a single arc), transform it into the basis of the reference coordinate system. The rest of Q_r is then transformed relative to this basis and the location of each point is used as an index into the corresponding bin in the hash table, where the basis and model number are added. In the recognition phase an arbitrary geometric feature in P_s is chosen and set as the basis. The rest of P_s is then transformed relative to it and each point 'votes' for its corresponding bin. The bins are histogrammed and the models above a threshold of votes are selected. The alignment method is then performed on these potential matches only. However, local deformations in Q_r and P_s will cause incorrect binning so robust techniques weight each vote over a neighborhood of bins. Another issue is the index voting distribution. For an efficient hash table each bin should have an equal probability of a vote.

HERMeS [27] uses a geometric hashing descriptor in a production system pattern matcher. Pànek and Vohradský [71] recognize that due to local distortions all reference spots in the neighborhood of a sample spot are candidates for a match. Each spot is matched only if both a neighborhood similarity measure and a geometric hashing method agree. The similarity measure compares the candidate match vector spots on the source image C_s and on the transformed image C_t with those of nearby landmarks L_1 to L_n:

$$\text{sim}(M) = w_d|C_t| + w_\alpha\left(\frac{\nabla\alpha}{\overline{\nabla\alpha}}\right) + w_\lambda\left(\frac{\nabla\lambda}{\overline{\nabla\lambda}}\right) \tag{33}$$

$$\text{where} \quad \nabla\alpha = \sum_{i=1}^{n} \frac{\alpha(C_s) - \alpha(L_i)}{d_i} \qquad \nabla\lambda = \sum_{i=1}^{n} \frac{|C_s| - |L_i|}{d_i}$$

$\overline{\nabla\alpha}$ and $\overline{\nabla\lambda}$ are the means of all candidates
$w_d + w_a + w_\gamma = 1 =$ weights

α is the angle function and d the distance between the landmark and the candidate spot. A geometric hashing method only accounting for translation was performed simultaneously. The space around a spot is quantized by both radial distance (polar coordinates) and absolute angle into 20–40 hash table entries. Local distortions were accounted for by using a special indexing technique in which the bit pattern of the indices have more identical bits the closer the bins are. In the recognition phase the confidence of a match is rated by the number of identical bits in its descriptor. In the case of one spot matching to multiple, the match with the highest confidence was assigned and the rest set unmatched. High confidence matches become new landmarks and the sample image is then re-registered. The best global match converges in a couple of iterations and empirically has a best case match efficiency and accuracy of 98%.

The CAROL system [81–83, 59] reduces the search space of the alignment method with a local search strategy based on the extended history of the Delaunay Net. False matches are removed through consistency checks between subsets. The positions of subset matches are then used to transform the sample image before a global matching takes place.

The warping on I_s is performed on a 5×5 regular grid. For each of the 25 subimages the 12 most intensive spots are selected for local pattern matching, outputting 25 lists of proposed matches. One match is selected from each based on the consistency of their transformations. These 300 spot pairs are used as landmarks to warp the image with a piecewise affine mapping m. A global matching process then starts in which every spot in P_r is matched with its nearest neighbor in $P_s \circ \mathrm{m}$, as long as their spot neighborhoods are similar. If for a grid square few neighborhoods can be matched the algorithm backtracks and chooses an alternative local matching and m. This is especially useful when considering multiple candidates for ellipse covering of complex regions [65].

The alignment method is only used in the local matching phase as only rigid transformations can be accounted for [81]. The 12 points are annotated with a discrete intensity score between 0 and 10. The angle between candidate arc matches must be less than α, the spot intensities within two units and for their lengths:

$$1 - \lambda \leq \frac{|\overrightarrow{ab}|}{|\overrightarrow{cd}|} \leq 1 + \lambda \tag{34}$$

The Hausdorff distance will therefore always be bounded above by:

$$(\lambda|\overrightarrow{ab}|) \tag{35}$$

so instead the optimal match set is the one with the largest cardinality. Candidate transforms are only generated on (λ,α)-similar arcs and all arcs in a match must be (λ,α)-similar.

The number of arcs tested is reduced further by using only those in the extended history (*Hist**) of the DN for P_s and P_r. CAROL builds up the DN incrementally starting with the most intensive spot. Each new Delaunay edge is added to *Hist**, together with the 'flipped diagonals' – all edges connecting the new point with opposite points in its neighboring triangles. Incorporating the flipped diagonals increases the probability of including edges from P_T whilst the size of *Hist** is still bounded linearly by 12*n*. *Hist** is also well suited to recognizing local patterns in images with up to a third of noisy spots. Geometric hashing is also used, on the principle that the arcs of the local patterns will be matched closely together on the target point set. The displacement vectors between the midpoints of each scaled sample edge and target edge are hashed. Rather than bin a histogram, they update the four corner nodes of the corresponding cube in a regularly spaced 3-D finite element grid, where the third dimension is the logarithm of the scaling factor. This outputs the list of the centroids and scalings of possible matchings ranked by their node scores.

Only the RPM, methods explicitly transform the image to account for distortions. Robust point matching methods estimate the point correspondence and transformation simultaneously and iteratively. Unfortunately they tend to be rather less intuitive than other methods.

Akutsu *et al.* [75] matched Restricted Landmark Genomic Scanning (RLGS) gels by extending the P time one dimensional dynamic programming method to two dimensions. In one dimension an iterative scheme can compute the maximum cardinality match by mapping the sample point set onto the reference point set in every permutation conserving the ordering, and choosing the mapping with minimum error. For two dimensions they use the heuristic of matching only the longest diagonal path of points in L_1 norm space. The sample image is then affine transformed with the least squares fit. The algorithm then interleaves ICP and 'local' transformations to match the rest of the spots.

Pedersen and Ersbøll [72] extended the RPM method of Gold *et al.* [84] to make it more robust to outliers and handle unequal point sets. The algorithm estimates the parameters of each patch of a piecewise affine transformation by solving a series of assignment problems by deterministic annealing (DA). Gold *et al.* [84] defined the doubly stochastic match matrix Ξ, which is a table of reference spots against sample spots where each element can take a value between 0 and 1 where 1 is a certain match. An extra row and column are added to Ξ to denote spurious or missing spots. All rows and columns must sum to 1. The objective is to find the affine transform $A + t$ and corresponding Ξ which minimizes:

$$E(\Xi, t, A) = \sum_{i=1}^{n}\sum_{j=1}^{m} \Xi_{i,j} ||P_r - t - AP_s||^2 + \\ + g(A) - \alpha \sum_{i=1}^{n}\sum_{j=1}^{m} \Xi_{i,j} \qquad (36)$$

$$g(A) = \gamma(a^2 + b^2 - c^2)$$

where

$$A = \begin{pmatrix} e^a & 0 \\ 0 & e^a \end{pmatrix} R(\Theta) \begin{pmatrix} e^b & 0 \\ 0 & e^b \end{pmatrix} \begin{pmatrix} \cosh(c) & \sinh(c) \\ \sinh(c) & \cosh(c) \end{pmatrix}$$

R(Θ) is the standard rotation matrix. α acts as a threshold error distance, and γ in *g(A)* regularizes the affine transformation by penalizing large scale and shear components.

In deterministic annealing (DA) optimization the objective function is systematically varied by introducing a temperature parameter *T*. DA essentially starts with a (trivial) convex optimization problem at a large temperature, and then repeatedly finds more accurate solutions by lowering it. An example of a single constraint DA problem is the winner takes all (WTA), to assign a 1 in Ξ for the largest number in a vector *Q,* and 0 for the others. This can be formulated continuously with a control parameter β (inverse temperature):

$$\Xi_i = \frac{\exp(\frac{1}{T} q_i)}{\sum_{j=1}^{n} \exp(\frac{1}{T} q_j)} \qquad (37)$$

The exponentiation ensures all elements of *M* are positive. As $T \to 0$ the maximum q_i tends Ξ_i to 1 and the rest to 0. However, our problem is a two-way WTA assignment, in that there is a further constraint that both the rows and columns in Ξ must sum to 1. The same procedure is employed but Sinkhorn's method must be used to normalize Ξ between iterations. Sinkhorn proved that a doubly stochastic matrix is obtained from any square matrix with positive entries by the iterative process of alternative row and column normalizations. Non-square matrices give a similar result.

The full point matching problem can now be derived. For fixed $A + t$ it is simply an assignment problem where:

$$Q_{i,j} = -\left(||P_r(i) - t - AP_s(j)||^2 + \alpha \right) = \frac{-\partial E}{\partial \Xi_{i,j}} \qquad (38)$$

So after each normalization the $A + t$ parameters are reestimated by a one step coordinate descent. Pedersen and Ersbøll [72] start from a high α and decrease it as *T* decreases to improve the robustness at the start of the process where pose and correspondence are uncertain. Their algorithm works separately on an extended area around and inside each patch in a piecewise affine transformation. The match matrices are then transferred to a global match matrix, where points processed multiple times enter a voting scheme for the most likely match.

3.1.3 Hybrid and direct methods

Complex spot matching traditionally follows simple landmark warping, but some researchers [41, 85] believe that the inaccuracies of the feature based approach are based on insufficient data. If one bases the automatic registration on the raw pixel values numerous features such as spot shape, streaks, smears, spot tails and background structure are available which are otherwise lost in the spot detection phase. The intention is a shift in complexity to the image warping stage.

Conradsen and Pedersen were ahead of their time, introducing direct registration to 2-DE in 1992 [53], but the specialist system required for speed made it unpopular then – one gel pair took a few hours on a GOP-302 image processor. A bicubic subsampling multiresolution approach removes course global deformations at lower resolutions before removing finer local deformations at higher resolutions. Starting at 64 × 64 resolution the images were convoluted with a 5 × 5 high pass filter. The minimum 3 × 3 mean squared error (MSE) between each sample pixel and the reference pixels in a 5 × 5 window is calculated and the match vector estimated by parabola interpolation between this MSE and the MSEs of the window's top left and bottom right. The resulting match vector field is smoothed by a 5 × 5 median filter and then warps the sample image of double the resolution. Rotation and scaling are determined by the match vectors in the centers of the four image quadrants. The algorithm iterates until 512 × 512 resolution. This method can only cope with deformations of up to 5% of the gel size.

It was not until recently that the processing power became cheap enough for direct registration methods. The Z3 system [41] is a hybrid feature based/direct registration strategy whereby the images are covered by small rectangles each containing a cluster of spots. As each sample rectangle is matched to the best reference candidate, a function constraining candidate selection is updated. This function defines a global transformation which increases in complexity as more rectangles are matched. Once all rectangles have been processed the sample image is warped by the final transform. The algorithm repeats on the transformed image until no further warping occurs.

Firstly, a sequence of covering rectangles (SCR) for the sample image is generated. The size of rectangle is selected to be large enough to differentiate them but small enough so that distortions can be approximated by translation only. Empirically, the spots are grouped into clusters of four and containing rectangles computed for each. These rectangles are processed in order of increasing distance from a high scoring (based on area and spot intensity) rectangle near the center of the gel. A transformation function for the sample image is refined as the rectangles are processed, initially based on the identity transform. The first rectangle's match vector defines a global translation, and as more match vectors are added the mapping becomes rigid body, affine and finally a Delaunay transformation, a piecewise bilinear mapping on the DN. In turn, the transformation function constrains the area searched in the reference image by the remaining matches to a bounding box centered on the transformed sample rectangle. The size of the bounding box varies depending on the statistical error of the previous match vectors. All possible rectangles in the bounding box are analyzed with a granularity of step size *s*, which is set as the width of the smallest visible spot.

The pixel level similarity measure used is a weighted combination of three factors tailored to 2-D gels: (a) The scalar product of the two gradient vectors, so that spots with similar shapes score higher. (b) A bonus value when both sample and reference pixels do not come from the background. (c) A bonus value when both represent intense spots.

Each match vector is refined by locally registering the matching sample and reference rectangles with optical flow techniques [86]. One of the first ten rectangles must be matched against a rectangle in the center of the reference. This 'seed' matching is vital to the rest of the matching and limits the global translation tolerated. Global rotation and scaling must also not exceed 10% and 5% respectively. Landmarks are necessary if this distortion is exceeded. Once all the match vectors have been calculated they are consistency checked to remove erroneous vectors. However a cluster of such vectors supporting each other may represent a tear in the gel instead.

In 2001 Veeser, Veeser *et al.* [85], created Multiresolution Image Registration (MIR), a direct registration algorithm with no recourse to detecting spots. They optimized the control points of a piecewise bilinear mapping with a quasi-Newton rooting finding technique and the closed form derivative of a cross-correlation similarity measure. Convergence occurs because the gel images have smooth gradients and therefore the cross-correlation with respect to the control points will be smooth and continuous. To avoid convergence to a local minimum, a multiresolution approach is used. Since the intention is for MIR to account for all global and local distortions, nearest neighbor or any spot matching algorithm to hand can be used to pair the spots.

Firstly, the gels are registered to a global rigid body deformation (scaling and rotation) by finding a mapping with the highest cross-correlation on heavily Gaussian subsampled images (32 × 32 pixels). The cross-correlation similarity measure is given by:

$$\mathrm{corr}(I_r, I_s) = \frac{\sigma(I_r, I_s)}{\sigma(I_r)\sigma(I_s)} \tag{39}$$

$$\text{where} \quad \sigma(I_r, I_s) = \frac{1}{|D|}\sum_{(x,y)\in D}\left(\left(I_r(x,y) - \bar{I}_r\right)\left(I_s(x,y) - \bar{I}_s\right)\right)$$

$\sigma(I_1,I_2)$ is the covariance between I_1 and I_2 and D is the domain of points considered for the registration process. A brute force optimizer is used on a limited space of admissible rigid mappings. In the second phase the corners of the 32 × 32 sample image are optimized as the control points of a bilinear mapping. The Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimizer [87] used does not require second derivatives and always moves in an uphill direction. In order to use this approach Veesar *et al.* [85] found the closed partial first derivatives of the cross-correlation with respect to the parameters Θ of transformation mapping *m*:

$$\nabla\mathrm{corr}(I_r, I_s) = \frac{\nabla\sigma(I_r, I_t)\sigma(I_t) - \nabla\sigma(I_t)\sigma(I_r, I_t)}{\sigma(I_r)\sigma(I_t)^2} \tag{40}$$

$$\nabla\sigma(I_t) = \frac{\sigma(I_t, \nabla I_t)}{\sigma(I_t)} \qquad \nabla I_t = \nabla m_\Theta \cdot \frac{\partial I_t}{\partial m_\Theta}$$

$$\text{where} \quad \nabla\sigma(I_r, I_t) = \sigma(I_r, \nabla I_t) \qquad I_t = I_s \circ m_\Theta$$

Once the course deformations have been found the algorithm moves up the Gaussian pyramid – optimization is performed on the gel images from 64 × 64 pixels to 512 × 512 and the piecewise bilinear mapping from 2 × 2 pieces to 16 × 16. Therefore at each stage finer and finer deformations are accounted for, up to one-quarter of the side length between control points. Furthermore, the control point optimizations are performed in parallel ('decoupled') for additional performance gains. Fig. 6 shows MIR in the authors' *proTurbo* analysis framework.

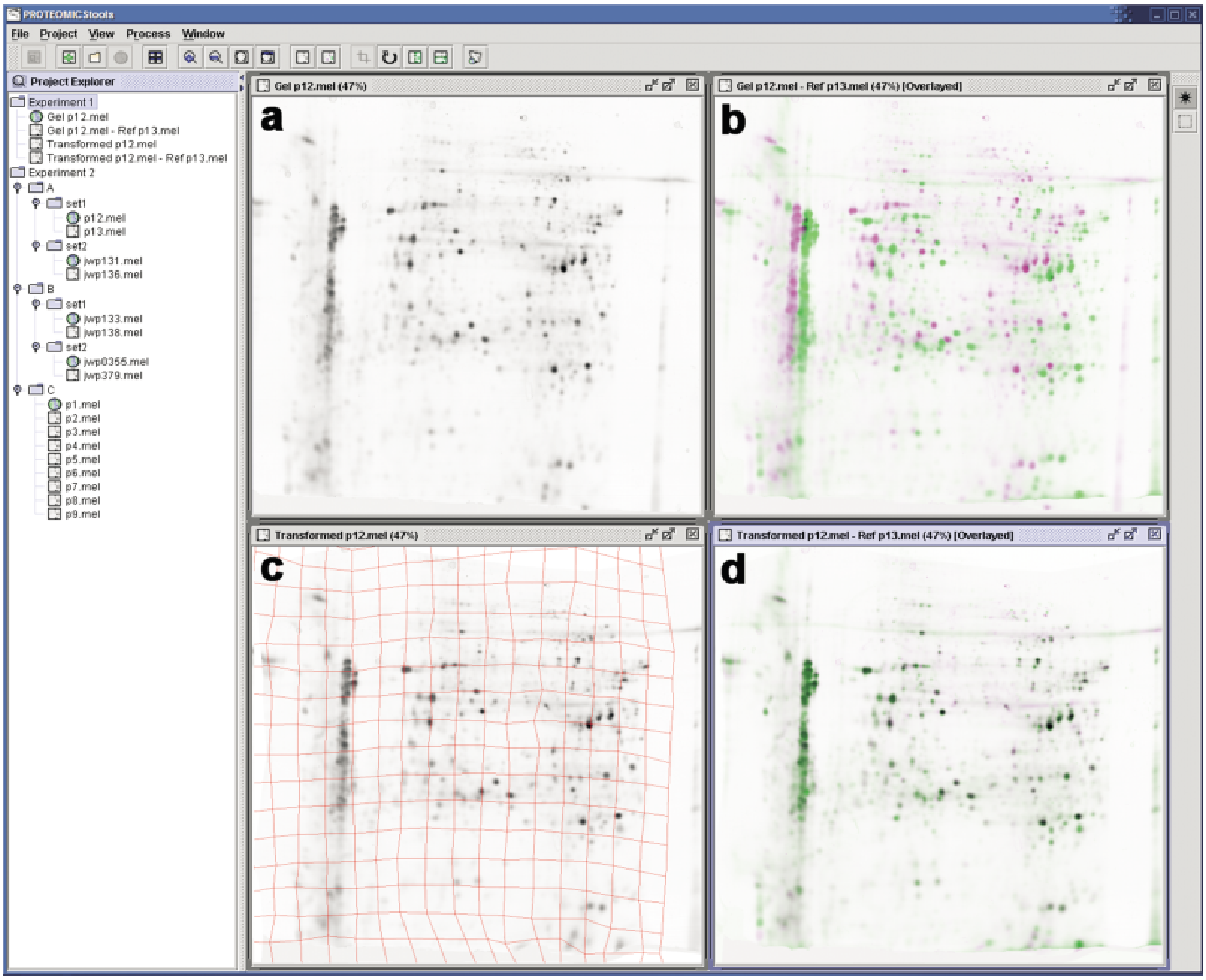

Figure 6. MIR image registration of Veeser *et al* [85] in Dowsey *et al.* image analysis toolkit *proTurbo* (http://vip.doc.ic.ac.uk/proturbo/). (a) The sample gel; (b) the sample gel (magenta) overlaid with the reference gel (green); (c) the registered sample gel with the piecewise bilinear mapping grid illustrated; (d) the transformed sample gel overlaid with the reference gel.

In a comparative study between MIR and Z3, MIR had equivalent speed and scored better 29 out of 30 times in a test where an expert quantified the mis-registered spots. On average all but 1% of spots could not be matched. *proTurbo*, sample data sets and MIR literature are all available from http://vip.doc.ic.ac.uk/proturbo/.

An alternative similarity measure, mutual information, has the advantage that it can be used for intra-modal registration, *i.e.* it does not require a direct relationship between the intensities of the reference and source images so it can be used to register images obtained by, *e.g.* different staining methods. Mutual information is given by:

$$\mathrm{mut}(I_r, I_s) = -\sum_{i \in H,} \sum_{j \in H,} P(i,j) \log_2 \left(\frac{P(i,j)}{P_r(i) P_s(i)} \right) \tag{41}$$

where H_r and H_s are histograms of the intensities in I_r and I_s respectively, *P*(*i*) is the probability of *i* in *H* and *P*(*i,j*) the joint probability of *i* in H_r and *j* in H_s.

In order to make the derivatives smooth enough for a quasi-Newton optimizer to converge, for sparse histograms we must distribute each bin entry over many neighboring bins. A Parzen window allows us to do this with some well known distribution *N* such as a Gaussian. The Parzen estimate of *P* is:

$$\tilde{p}_N(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{w\left(\frac{x - x_i}{\varepsilon(N)}\right)}{\varepsilon(N)} \tag{42}$$

where ε is a strictly positive scaling factor that controls the width of the Parzen window function *w*. Baker *et al.* [38] calculate the first derivatives by a difference method sampling two close points, and use a multiresolution piecewise bilinear mapping. Thévenaz and Unser [88] developed closed form first and second derivatives of the mutual information function given a B-Spline basis function Parzen window, as a Gaussian Parzen window does not satisfy the partition of unity constraint and therefore will be affected by the transformation model used in the registration procedure. They use the Marquardt-Levenberg optimization strategy rather than the BFGS.

Recently Gustafsson *et al.* [46] added a transformation smoothness constraint to a log-likelihood similarity measure. Their method is otherwise similar to MIR except that: the first partial derivatives are calculated using difference methods; the optimizer used is a conjugated gradient algorithm from MATLAB (The MathWorks, Natick, MA, USA); the registration is performed on the log images *L* = log(*I*) to compress the data to one order of magnitude. If we assume the image intensities after warping follow a Gaussian distribution:

$$\mathrm{L_s} \circ \mathrm{m}_\Theta = \mathrm{L_t} \sim \mathrm{N}(\mu_\Theta, \sigma^2) \tag{43}$$

then, disregarding additive and scaling constants, the penalized log-likelihood similarity measure is a probability *P*:

$$P(L_t | L_r, \Theta) = -\frac{1}{\mathrm{card}(L_t)} \sum_{(x,y) \in t} (L_t(x,y) - \mu_\Theta)^2 - \lambda D(\Theta) \tag{44}$$

where *D* is a smoothness constraint – the bending energy of thin plate splines in a finite window:

$$P(\Theta) = \sum_{i=1}^{2} \sum_{j=1}^{2} \sum_{k=1}^{2} \int_{(x,y) \in \mathrm{domain}(L)} \left(\frac{\partial \Theta_i}{\partial x_j \partial y_k} \right)^2 d(x,y) \tag{45}$$

3.1.4 Intensity bias correction

No current software package attempts to correct the gels for pipetting, gel focusing and staining errors, relying instead on extrapolating the background variability to the spot areas. Regional expression variation is essentially a regional variation in mean intensity over the image. Similar problems have been reported in the areas of magnetic resonance imaging and microscopy. In Magnetic Resonance Imaging (MRI) 'gain field correction' is the process of removing intrascan intensity variations due to inhomogeneity in the B_0 (constant magnetic) and RF (radiofrequency) excitation fields, and from regional differences in the magnetic properties of the tissues [89]. In microscopy the same problem crops up due to variable slice thickness or nonplanar surfaces [90] and its resolution is called 'shading correction'. In both sciences removing these inhomogeneities is essential for the post-segmentation of the image into tissue classes.

The multiplicative gain field in MRI has been analyzed to be smoothly varying, and therefore separable from the underlying tissue structure, which is assumed to contain a small number of tissue classes each with approximately constant intensity values. Little work however has been performed in 2-DE to analyze the properties of staining since Rabilloud [5]. If we assume the inhomogeneities are very slowly varying over the gel (so that regions appear homogeneous) then they can be corrected without affecting true differential protein expression, which causes far more abrupt changes in intensity. The gain field found is relative to a reference gel as otherwise one would have to make assumptions about the protein expression distribution over a single gel. The distribution must have a small variance, but for gels with few spots the variance is likely to be unfavorably high. Mathematically the gain in expression intensity β can be expressed as:

$$I_b(x, y) = I_o(x, y) \beta(x, y) + n(x, y) \tag{46}$$

where I_b is the measured image, I_o is the image representing true protein expression and *n* is image noise. If one prefilters to remove *n* we can take logs to turn β into an additive bias field component:

$$\log(I_b(x,y)) = \log(I_o(x,y)) + \log(\beta(x,y)) \tag{47}$$

In MRI and microscopy many retrospective techniques have been applied, either by incorporating the statistical properties of the underlying tissue classes, or not. They include homomorphic filtering and homomorphic unsharp masking (HUM) [91], information minimization [92], surface fitting [93], overlapping mosaics [94], fuzzy c-means and expectation-maximization [95].

Lai and Fang [93] developed a shape-from-orientation approach that could be incorporated into the traditional spot detection and matching method for 2-D gel analysis. They formulate the gain field correction problem as a surface fitting from sparse orientation constraints in a regularization framework. It is much easier to determine whether the pixels in a small neighborhood belong to the same tissue class. Therefore, assuming each tissue class has constant intensity, the orientation (gradient) is sampled at many locations over the log image. Recovering the gain field function from these sparse orientation constraints is ill-posed, but can be reconstructed by integrating these constraints in a regularization framework. The regularization formulation finds the I_b that minimizes the following thin plate spline energy function (a constraint on the smoothness of the bias field):

$$U(\beta') = \iint_{I_x\, I_y} \alpha(x,y) \begin{bmatrix} \left(\dfrac{\partial\beta'(x,y)}{\partial x} - \dfrac{\partial I'(x,y)}{\partial x}\right)^2 + \\ \left(\dfrac{\partial\beta'(x,y)}{\partial y} - \dfrac{\partial I'(x,y)}{\partial y}\right)^2 \end{bmatrix}$$

$$+\lambda\left[\left(\frac{\partial^2\beta'(x,y)}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2\beta'(x,y)}{\partial x\partial y}\right)^2 + \left(\frac{\partial^2\beta'(x,y)}{\partial y^2}\right)^2\right]dxdy \tag{48}$$

where $\beta'(x,y) = \log(\beta(x,y)) \qquad I'(x,y) = \log(I(x,y))$

α is 1 in smooth regions and 0 in high gradient areas; λ is the regularization parameter controlling the degree of smoothness. Lai and Fang discretize the bias field and energy function into a nonconforming finite element grid with second order polynomial interpolation between elements and search for a solution using a preconditioned conjugated gradient algorithm. For 2-D gels the orientation gradients would be the ratio in expression between the two spots in each pair-wise match. Figure 7 shows the Lai and Fang approach performed between a reference and a sample 2-D gel. This approach is disadvantaged by the inherent uncertainty in expression values computed by spot detection techniques. Using the full image information instead would lead to a more robust approach.

Expectation-maximization (EM) is a Bayesian approach often used in estimation problems when some of the data is missing, in the MRI gain field correction case this

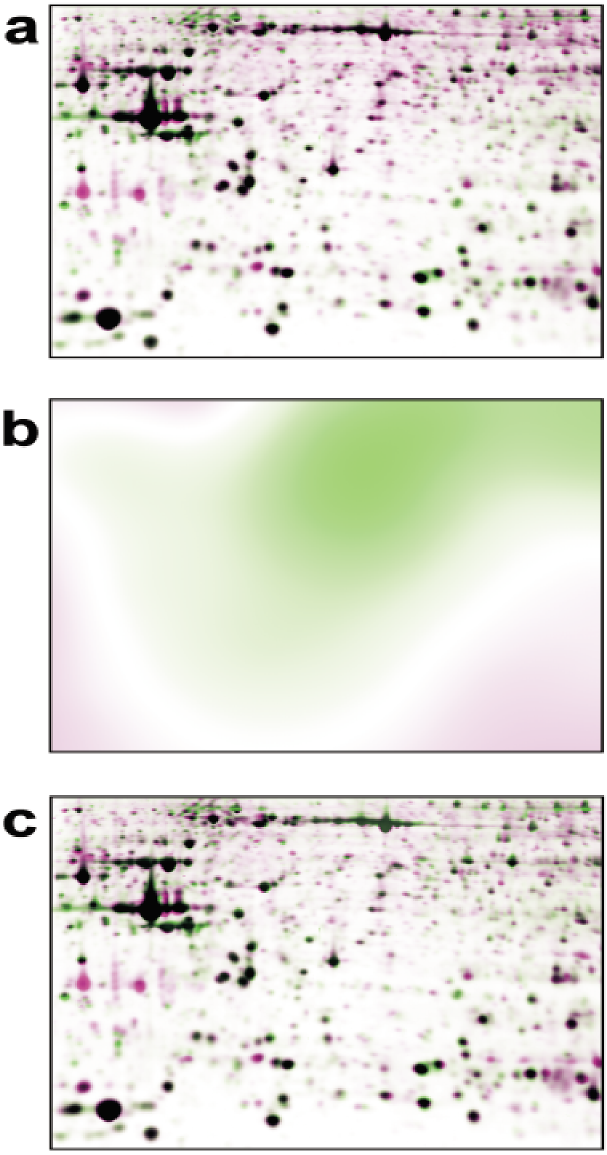

Figure 7. Lai and Fang's [93] shape-from-orientation bias field correction approach applied to find the relative bias between a reference (green) and a sample (magenta) 2D gel. (a) The original gels overlaid. (b) The thin plate spline bias field computed from the ratio of spot volumes between the two gels, given as a log additive field. Magenta represents where the reference spot volumes are greater than the sample spot volumes, and green represents where the sample volumes are greater than the reference spots. (c) The normalized reference (multiplied by the bias field) overlaid with the sample. Notice that spot volume takes into account the size of the spot whilst the bias field is purely spatially independent. So the bias correction can only increase or decrease the intensity of a reference spot to compensate for a bigger or smaller sample spot. Normalized spots overlaid will therefore have magenta or green fringes but slightly green or magenta centers to compensate. The following differential expression analysis must allow for this phenomenon.

is the tissue class segmentation. MR images have a small number of classes and a large difference in intensity is assumed between them. This is modelled by a Gaussian finite mixture model, where the intensity in each class follows a Gaussian distribution with small variance. The EM algorithm iteratively alternates evaluations of two expressions [96]: (a) Estimation: calculating the posterior tissue class probabilities when the bias field is known; (b) Maximization: maximum-*a-posteriori* (MAP) estimator of the bias field when the tissue class probabilities are known.

In the initial loop the bias field is assumed to be zero. The actual bias field and tissue class probabilities converge in 5–10 iterations. Let n be the number of pixels in the image. Then the smoothness criterion of the bias field is modelled by an n-dimensional zero-mean Gaussian prior probability density with an $n \times n$ covariance matrix. This matrix is impractically large to manipulate unless we represent it as LL^{T}, where L is a low-pass finite impulse response filter. This has the effect of reducing the probability of higher frequency bias field components, which is the desired effect anyway.

It is suggested that 2-DE images are modelled as a mixture of a Gaussian background component and a new model for the spots. Regional spot intensities are adjusted to the model's homogenous intensity distribution within a smoothness constraint. Advantageously, the background subtraction and gain field correction are performed in parallel. The Bayesian approach is notable in that it outputs probabilities for the classes and bias field so these confidence intervals can be reflected upon further down the processing pipeline.

3.2 Differential expression analysis

One important goal of proteomics is to find the protein expression that is induced (new spots), inhibited (missing spots) or changed (differences in spot intensity) between biological samples taken under differing conditions, *e.g.* diseased, treated or from a different individual. Since the spot detection, quantification and matching processes do not have total accuracy, often the task has been to compare a set of reference gels with a set of sample gels to statistically separate true differential expression from noise.

The intermixed reactions cells have to altered conditions suggest that large scale analyses over a large set of different patients and/or conditions could find intrinsic trends in the differential expression. For example, a set of proteins may be interrelated through a pathway influenced by another set of proteins, and when affected by a condition their expression may be correlated. Once the mutations are found classification analyses attempt to find these relationships. Utilizing prior results, classification also becomes a means to identify the unknown origin of tissue samples or the progression of a disease or treatment state [97] – unique proteins or protein patterns can be used as markers for subsequent identification purposes [98].

Two different methodologies can be identified, whether we characterize a maximal number of spots regardless of their possible role in cell life, or choose a set which are likely to fall into a functional group showing model behavior [99]. The power of the model approach lies in its ability to detect expected protein regulation in time course experiments. In time course experiments gels are produced for biopsies at regular points in the developmental cycle.

3.2.1 Mutation detection

Mutation detection is based on comparing the set of n reference spot lists $R_1 \ldots R_{\mathrm{n}}$ with the set of m sample spot lists $S_1 \ldots S_{\mathrm{m}}$. The spot lists are a set of tuples $\Theta\epsilon\Theta$ defining the identity, position, intensity, and other parameters of the spot model used in detection. The same protein in each list is assigned the same identifier. A tuple of spots with the same identity is called a spot profile. Correlating the identifiers is achieved with a graph covering of the pair-wise match lists. A 'graph covering' states that every spot in each spot list must be related to its counterparts in the other spot lists through a path of match lists. Due to missing spots, it is quite possible that unless every pair-wise permutation of spot lists are matched, matches will be missed. However, DNAInsight [100] only requires each spot list to be matched to a global reference. The set of gel images are then warped to the reference image by a Delaunay transformation. Unmatched spots are then clustered (see Section 3.2.2) to identify new transitive matches. Missing landmark spots can be extrapolated in a similar way [101].

Early univariate methods of mutation detection include student's t-test [34], the Mann-Whitney test and Analysis of Variance (ANOVA). The t-test is a statistic for measuring the significance of the difference in means between two Gaussian distributions. For each spot the mean reference %IOD Eq. (22) and mean sample %IOD are compared, and if they found to be different above a user set confidence level, differential expression is flagged. *%IOD* must be used to compensate for global staining and loading anomalies. The Mann-Whitney test is a non-parametric alternative that does not need to assume Gaussian distributions, but is based on ranking the spot ranks rather than on exact volume data. One-way ANOVA [102] is the *t*-test generalized to comparing more than two population means, and is useful for mutation detection over many treatments [103]. The populations are assumed to be Gaussian distributions with equal variance. *n*-way ANOVA can be used when there are more than one factor affecting the populations, *e.g.* treatment, patient, batch. It can find significant mean differences with respect to each factor ('main effects'), or if there is significant dependency between specific factors analyzed ('interaction effects').

Taylor and Giometti [104] observed that protein spots in a 2-D pattern, due to regional loading and staining anomalies, often represent correlated rather than independent measurements. They use PCA as an implicit bias field correction to compensate for this variation. The set of reference spot lists is used as training data. The PCA is run on the IOD data of the spot profiles. Data points with missing spots must be removed. After processing the first x principle components contain the most common volume changes over the set whilst the other $n - x$ are dominated by single spots and therefore indicate noise. Intuitively, the first principle component is interpreted as the global staining and loading error between the gels. The variance between each spot's IOD and its reprojection from the first x principle components is calculated and stored. The sample spot lists are compared to these predicted values in the screening phase. Spots that deviate by more than three times their variance are flagged as candidate outliers. If the same spot is flagged multiple times it is flagged as true differential expression. In tests the overall efficiency of the algorithm was 75%, so therefore much user interaction is required to dismiss false positives. Also, new or missing spots cannot be detected by this method.

If a particular model of activity is known, constraint analysis can be used to search the spot lists for instances of the model. Post-translational modification chains of spots, putative point mutations and putative precursor product-pairs are examples of models that can be detected [101]. For example, say a small peptide is cleaved from a precursor yielding the product:

$$\text{precursor}_1 = (\text{signal}_2 + \text{product}_3) \tag{49}$$

Assuming no p*I* change. The following constraints constitute the model:

$$|\mathrm{p}I_1 - \mathrm{p}I_3| < \mathrm{T}_{\mathrm{p}I}$$

$$|M\mathrm{r}_1 - M_{\mathrm{r}3}| < \mathrm{T}_{M_\mathrm{r}} \tag{50}$$

$$\frac{(O_1 + o_3)_\mathrm{a}}{(o_1 + O_3)_\mathrm{b}} \sim 1.0$$

where T_{pI} and T_{M_r} are thresholds and O and o the protein concentrations for the two spots in gels *a* and *b*. Constraint analysis simply searches the gel for spots satisfying the model constraints.

Two sources of information are so far unused in mutation detection. Firstly, if the pair-wise statistical analysis were performed simultaneously on the whole set of reference and sample gels, then small insignificant expression changes over one pair could become significant when reinforced by the same small changes in the other pairs. Arguably, the sensitivity of the algorithm would then increase dramatically. Secondly, quantifying changes on the raw pixel values rather than the spot lists would remove the all too regular errors brought on by the spot detection and matching processes. Spot detection would then become a process of clustering groups of pixel variation and validating they belong to a spot or spot shoulder and are not just the signs of random noise. This method we call the statistical approach model (SAM).

3.2.2 Trend analysis

Each gel is characterized by a large number of parameters – its spot list. Since biological processes are indeterminate, it is a complex problem to determine all the component parts and explanations of all the interrelated relationships and functions [101]. A large set of gels can provide a powerful interference tool for comparisons of mechanisms in diverse systems. The *n* spots present in all *m* gels are represented as *n* data points in *m* dimensions as before. The aim is to reduce and transform the dimensionality so that distances indicate the strength of the relationship between spots. Typically this factional space is the first few eigenvectors of PCA [103, 105, 106] or correspondence analysis (CA) [97, 98, 107]. Nearby spots can then be classified into groups by cluster analysis.

CA is an exploratory technique related to PCA which finds a multidimensional representation of the association between the row and column categories of a two-way contingency matrix *F*. It finds scores for the row and column categories on a small number of dimensions which account for the greatest proportion of the χ^2 association between the row and columns, just as PCA accounts for maximum variance. CA can cope with nonmetric and nonlinear data, but only outputs descriptive rather than quantitative relationships. The χ^2 statistic matrix X is calculated from the contingency probability matrix P (each cell's frequency divided by the total frequency):

$$\chi_{\mathrm{ij}} = \frac{O_{\mathrm{ij}} - E_{\mathrm{ij}}}{\sqrt{E_{\mathrm{ij}}}} = \sqrt{f_{++}}\left(\frac{p_{\mathrm{ij}} - p_{\mathrm{i}+}p_{+\mathrm{j}}}{\sqrt{p_{\mathrm{i}+}p_{+\mathrm{j}}}}\right) \tag{51}$$

where O is the observed and E the expected frequencies, f_{++} the total frequency, p_{i+} the marginal row probability and p_{+j} the marginal column probability. The $\sqrt{f_{++}}$ term is removed to scale the eigenvalues to ≤ 1, and then the eigenvectors are calculated. Interpretation then follows as in PCA.

In 2-DE CA is performed with spots as row, gels as columns and IOD Eq. (22) values as cells in the contingency matrix. In factorial space each gel is the center of mass for its spots and each spot is the center of mass for its gels. The relative positions indicate which spots characterize each gel the best. The contribution a spot gives to a factorial axis indicates its significance.

Cluster analysis is a technique which forms groups of related entities in a data set. Ideally the aggregation criterion minimizes the centered intra-class seconds moment whilst maximizing the inter-class moment. The similarity metric usually takes the form of a distance metric in factorial space, either Euclidian, Mahalanobis (Euclidian distance weighted by the variance in each dimension) or Haussdorff. If the number of clusters k is known, the K-means clustering technique used in [105] first randomly assigns each entity to a cluster and calculates the cluster centroids. The entities are then iteratively assigned to the cluster with nearest centroid, and the centroids recalculated, and so on until no reassignments occur. If k is unknown, the agglomerative hierarchical technique can be used. It starts off each spot in its own group and then merges the most similar groups, and so on. The output is a tree, a dendrogram, where the further to the leaf nodes you get the closer in similarity are the children. The cluster algorithm in [98, 103] is the unweighted pair group method with arithmetic mean (UPGMA), which is a simple heuristic algorithm. Basically, the algorithm iteratively joins the two nearest clusters until one cluster is left. UPGMA implicitly assumes the mutation probability for each spot is the same, which can lead to misleading results.

ChiClust [108] uses clustering to extract statistically significant protein expression patterns from a large set of gels. From this the ChiMap tool calculates and displays the differential expression. Ward's minimum variance method [109] is used. In Ward's method the sum of the squared distances from each member to the group centroid is used as an indicator of dispersion. Groups are joined only if the increase in this sum is less for that pair of groups than for any other. The distance metric in ChiClust is the ratio in IOD between matched spots:

$$\psi(m) = \frac{\%\mathrm{VOL}(r) - \%\mathrm{VOL}(s)}{\max(\%\mathrm{VOL}(r),\ \%\mathrm{VOL}(s))} \times 100\% \qquad (52)$$

where $(r, s) = m \in M$. The overall homotopy for a match list is:

$$\%H(M) = \left[1 - \frac{1}{\mathrm{card}(M)}\sqrt{\left(\sum_{m\in M}\psi(m)\right)^2 - \sum_{m\in M}\psi(m)^2}\,\right] \times \times 100\% \qquad (53)$$

where the partners of unmatched spots are given zero volume. An in-house program 'Concatenate' creates a redundant graph covering of the match lists to consistency check each match. ChiClust then searches for patterns of similar or different homology between spot lists and clusters the remaining spot lists by their distance. ChiMap displays the percentage differential output Eq. (52) as a color coded graph.

Recently Jessen *et al.* [110] analyzed the applicability of partial least squares (PLS) regression [111], also similar to PCA but also incorporating prior expert knowledge of the samples to aid discrimination. Whilst PCA is based on the SVD of X^TX, where X is the data matrix, PLS equivalently works on X^TY, where Y contains binary indicator variables expressing the prior knowledge. The first principle component of PLS therefore represents the highest variation that correlates to the indicator variables. It comes as no surprise that if Y enumerates the spots present for each class then the clustering on the first two principle components has a higher success rate. However, the aim instead is to find the most significant spots, and so jack-knifing is used to find spots with significant (*e.g.* $>5\%$) regression coefficients. Jack-knifing summarizes the partial model perturbations caused by leaving, in turn, different samples out from the model parameter estimation. The PLS iteratively run on the significant spots as long as the model improves. The final spots are all highly related to the classification and, advantageously, spots that contribute only in combination with other spots are also present.

For time course experiments the previous methods do not offer any greater functionality. $S_{n,m}$ spot lists are available, with n runs and m time points, together with the graph covering set of match lists. There are therefore $n \times m$ time course spot lists.

Vohradský *et al.* [112] first compared cluster analysis and UPGMA cluster analysis to classify the stages of *Streptomyces coelicolor* growth, and then he compared spot profile classification [112], PCA, cluster analysis and artificial neural networks (ANN) [99] in the detection of transition phase. The dramatic difference in expression profiles in this period can be characterized by the sharp up-regulation of some proteins (T+) and down-regulation of others (T−). In spot profile classification the first and second derivatives in each spot profile were calculated and profiles with good quality extrema extracted. These became the search template for classification, and the other profiles were classified by their correlation with the template. This simple algorithm proved fairly powerful but relies on setting arbitrary thresholds. PCA and cluster analysis were run on the $n \times m$ dimensional space of spot lists, where n~3,4,5 and m = 16. PCA could only find T+, identified as the correlation between the third and fourth eigenvectors. Cluster analysis grouped T+ successfully but grouped T− into several clusters.

ANNs are based on the parallel architecture of animal brains. They are a collection of simple computation units called neurons connected by synapses where the scalar output of each neuron is a weighted sum of its scalar inputs (the 'transfer function'). With each consecutive input a 'perceptron' network adapts by altering its weights w_i by an amount proportional to the difference between the desired output d and the actual output a:

$$w_{\rm i} = w_{\rm i} + \varepsilon(d - a)i_{\rm i} \qquad (54)$$

where $i_{\rm i}$ is the input and ε the learning rate. A single perceptron can linearly separate the data space into two half-spaces. Back-propagated networks develop this notion further with extra hidden layers (layers additional to the input and output layers, not connected externally). The hidden layers learn to recode their inputs. The transfer function of the hidden neurons is log-sigmoid:

$$a = \frac{1}{1 + \exp\left(\sum_{\rm i=1}^{n} w_{\rm i}\, i_{\rm t}\right)} \qquad (55)$$

where n is the number of inputs. The network topology is constrained to be loop free ('feedforward'). Any nonlinearly separable mapping can be learned, given two hidden layers. However, with too few neurons the network cannot approximate the solution satisfactorily, and with too many the network will overfit the data – not generalize sufficiently. Training the network consists of a forward pass where the outputs and output unit errors are calculated, and then a backward pass where the output unit errors are used to alter the weights. This is repeated over and over again until the output neuron error is below a threshold.

Vohradský found through experimentation that m = 16 input nodes, 30 neurons in hidden layer one and three in hidden layer two was suitable for classification with T+ and T− templates. One output neuron gave the answer to whether the template matched or not. Two networks were trained, one for each template, by initially requesting the network output +1 for the template and −1 for a random selection of spot profiles. Scores above +0.9 were manually classified as +1 or −1 and the training repeated until no new correct profiles were found. Automatic training with an initial training set of the template contaminated by Gaussian noise gave similar results. His neural network method could classify spot profiles into T+, T-, and 'other' with few false positives when the training set was greater than 150 spot profiles. A further advantage is that new spot profiles can be classified without retraining the network.

A direct approach to time course studies is currently used in MRI. FAMIS (Factor Analysis of Medical Image Sequences) [113, 114] finds the static images that make up the image sequence in such a way that a linear combination of them can produce each image in the sequence. In other words, an image sequence S can be represented as the sum of K underlying images $I_{\rm k}$, each weighted by their kinetics over time $F_{\rm k}$ + an error term E:

$$S(p,t) = \sum_{\rm k=1}^{\rm K} I_{\rm k}(p) \cdot F_{\rm k}(t) + E(p,t) \qquad (56)$$

Statistically, the sum of $I \cdot F$ represents the common variance and E represents the specific variance of each image. There are infinite solutions to this equation so we usually place a positivity constraint on the factor images and kinetics. PCA is then used to find orthogonal factors. If oblique (*i.e.* dependent) factors are required, an optimization step is performed which maximizes the variance between them. However the problem with oblique factors is that their number must usually be prior knowledge. In 2-DE gels it is expected that the principle static image will be the static protein expression in the sequence and tertiary facts are the differential expression. The kinetics represent the intensity changes in the expression at each point in the sequence.

4 Data presentation

Although interpretation of the output from differential protein expression analysis using 2-DE depends ultimately on the rigorous statistical and other tools that are applied to the numerical quantitation data, it is very useful for the biologist to have access to simple methods for visual presentation of the data. Thus all 2-D gel analysis packages provide facilities for displaying individual 2-D gel images, and the 'master' or 'average' images for sets of 2-D gels. In the simplest implementation these whole gel images (or 'zoomed' regions of the protein pattern) are displayed in tiled windows. It can also be very useful if the 2-D gel images can be superimposed ('stacked') [73, 41] and then animated to display sequentially each gel in the data set. This is a very rapid and effective tool, provided that the 2-D gel images are well registered, for the user to detect visually spots that are differentially represented in the data set. Such superimposition tools are now implemented in most of the commercial software packages and are, of course, essential in packages offering image registration [85, 41]. This is also a fundamental component of the “Flicker” tool developed by Lemkin [39] for comparison of 2-D gel images across the internet [40].

In most commercial software package, graphs of bar charts are used to provide a representation of quantitative spot data and to reflect trends. These are most usually displayed as histograms or bar charts of spot intensity data for each 2-D gel in the data set. Alternatively, if the 2-D gels can be formed into replicate groups, then the average intensity values and standard deviations of the groups can be displayed. Such charts provide a useful tool to display quantitative trends between individuals or groups.

It is important that appropriate tools for data presentation are implemented during the development of new methods and techniques for the analysis and data-mining of 2-D gel data, *e.g.* cluster analysis, PCA, neural networks.

5 Large scale analysis and computational issues

The demands created by the need for storing and processing thousands of gels create performance and efficiency concerns. Cluster computing is perhaps the most cost effective and modular approach to harnessing increased computing power. Its convenience lies in its possibility of harvesting resources already available and unused in your organisation. Databases provide a standard means to store, query and retrieve gel and spot information. Research in this area has aimed at confirmative and federated databases so that independent laboratories can publish their data on the web in a uniform and fully interoperative fashion.

5.1 Cluster computing

In the 1980’s the general consensus was that performance was best improved by faster processors. This idea was challenged by the advent of parallel processing. Since 1990 there has been an increasing trend away from expensive parallel supercomputers towards interoperating networks of workstations (NOW) [115]. At a basic level a cluster is a collection of workstations or PCs that are interconnected *via* some network technology. Fig. 8(a) shows the performance gain a cluster has over a standalone workstation for the registration of a set of gels using the MIR algorithm [85]. It follows a lambda law, where the improvement reduces logarithmically as the cluster size increases. A cluster works as an integrated collection of resources, coordinated by fault tolerant Single System Image (SSI) middleware. In heterogeneous clusters all nodes can have different architectures and run different

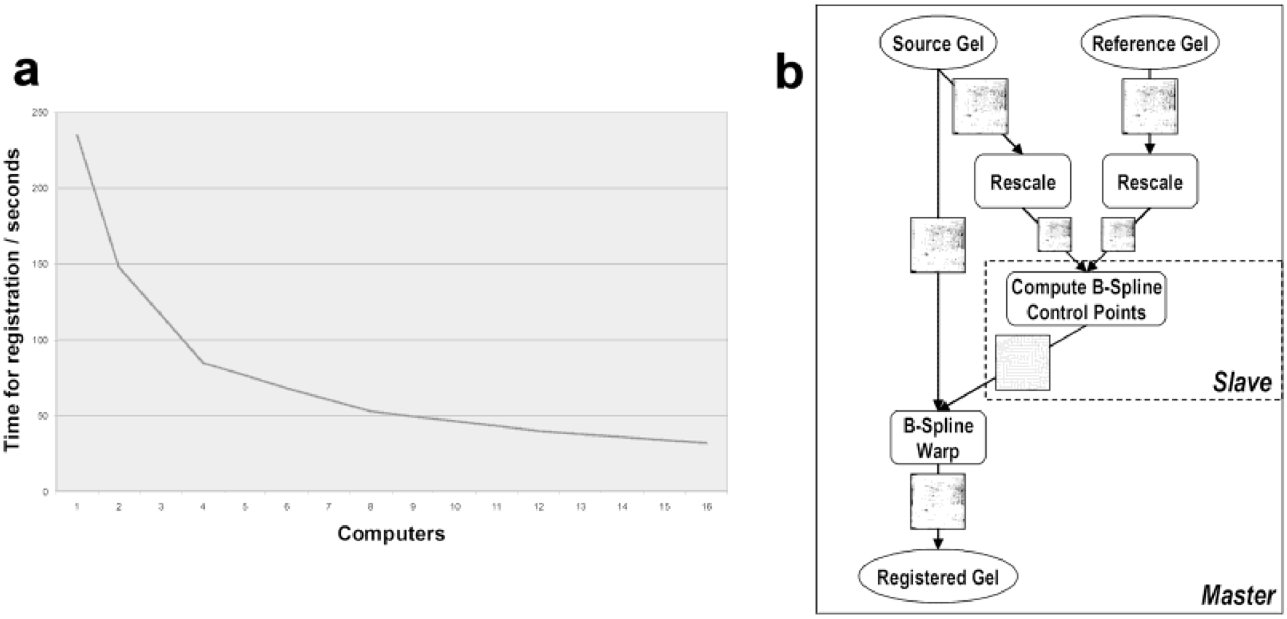

Figure 8. Cluster computation. (a) Time for MIR to register 32 gel pairs on a single AMD Athlon 1400 Mhz computer, and on clusters of size 2–16, using the master/slave paradigm. (b) DAG for MIR. The source and reference images are firstly rescaled on the master to optimize communications bandwidth. The slave calculates and sends back the piecewise bilinear mapping. The master then warps the original source image by the mapping.

operating systems. The computational grid technology of the Globus project [116] extends this resource sharing to dynamic collections of individuals and institutions over the internet, whilst Condor [115] is similar to a traditional batch queuing system but acts opportunistically stealing computing power. For instance, Condor can be configured to only use desktop machines where the keyboard and mouse are idle. Should Condor detect that a machine is no longer available (such as a key press detected), Condor is able to migrate the job to a different machine. Studies have shown that 60–90% may be idle at any given moment, which rises to almost 100% off-peak.

Resource management and sharing (RMS) [117] in the SSI dynamically considers several criteria, such as the response time, file accesses, communication between programs and memory needs in the cluster. A parallel program can be represented by a directed acyclic graph (DAG). Each node in the DAG represents a task. Each edge represents a data dependency. The nodes are weighted by their time to completion and the edges are weighted by the communication times between tasks. The precedence constraints of a DAG dictate that a node cannot start execution before it gathers all of the messages from its parent nodes. Figure 8(b) shows a coarse DAG for the MIR registration process. Data communication in the DAG method requires support from the applications programmer through message passing. Two such high-level message passing libraries are the Message Passing Interface (MPI Forum, http://www.mpi-forum.org/), and the Parallel Virtual Machine (PVM) [118] from Oak Ridge National Laboratory (TN, USA). Condor-PVM [119] integrates Condor's resource management with PVM's heterogeneous fault-tolerant message passing [120].

When designing a parallel algorithm, Foster [121] suggests organizing the process into four stages: partitioning, communication, agglomeration and mapping. One paradigm, task farming, is suitable when independent jobs are replicated, and usually involves a master/worker environment where a single master coordinates data to and from the worker, who process the subtasks. In geometric decomposition the problem domain is broken up into smaller domains and each process executes the algorithm on each part of it. Task farming is relevant in the processing of large quantities of 2-D gels, and geometric decomposition is very relevant in image processing. For instance, You and Shen [122] partition each image into a grid of blocks and pipeline each block separately.

5.2 2-DE protein expression databases

A review of proteome databases including protein and nucleotide sequence databases, pattern, domain, structural and post-translational modification databases is beyond the scope of this review. For a discussion of this area, the interested reader is referred to [123] and to resources on the web, for example *via* the ExPASy molecular biology server [124] (http://www.expasy.org/; http://www.ebi.ac.uk/services/). More pertinent to the present review are databases that are based on 2-DE separations of a particular organism, tissue, cell or subfraction. These databases comprise images of the 2-D gel separations and associated text containing information such as sample type, methods used for the study, and so on. In some cases large numbers of proteins from gel spots have been identified, especially techniques of mass spectrometry have become the industry standard for protein identification and characterization. In these cases, there is then a considerable amount of textual information associated with each spot, which can often be accessed *via* a "clickable" hyperlink from the 2-D gel image. This information usually includes protein p*I* and M_r values, method of identification, tissue distribution, disease associations, and hyperlinks to other relevant datbases. The SWISS-2DPAGE [125] (http://www.expasy.org/ch2d/) was a pioneering venture in this field and continues to be one of the most comprehensive 2-DE databases containing protein maps for human, mouse, *Escherichia coli*, and other species. In addition there are an increasing number of 2-DE databases for a wide range of species including mammals, yeast, plants, bacteria, viruses and specific cell lines [126–129]. A comprehensive list of these databases is maintained at http://www.expasy.org/ch2d/2d-index.html.

5.3 Interfacing and integration

As discussed in Section 5.2, an increasing number of 2-DE protein expression databases are being established. While some of the commercial 2-D gel analysis software systems include their own internal databases, these are not constructed in a standardized way. Moreover, they are generally only accessible internally, are usually very system specific and do not allow the 2-D gel images and associated data to be shared between institutions. Currently this is best achieved using the WWW *via* the internet. In order for such databases to be of maximal use to the scientific community they should ideally be constructed to standards that will allow full integration of data between all proteomic databases [130]. Such standardisation is difficult to achieve, although some steps towards this goal are being made by the Proteomics Standards Initiative (PSI) organised by the Human Proteome Organisation (HUPO). The aim of PSI is to define community standards for data representation in proteomics, and to facilitate data comparison, exchange and verification. The current state of this initiative can be reviewed at http://psidev.sourceforge.net/. However, PSI is currently not at the stage of dealing with 2-DE protein expression data. In the meantime, it has been suggested that such 2-DE databases should be constructed according to a set of fundamental rules [131]. Databases conforming to these rules are said to be 'federated 2-DE databases'. A list of existing federated 2-DE databases, and other databases conforming to at least some the rules, is maintained at http://www.expasy.org/ch2d/2d-index.html. Currently the majority of 2-DE protein databases are constructed manually, but in future there will be a need to develop tools for automated database construction. The freely available software package Make2ddb has been designed for automatic construction of federated 2-DE databases [34, 132].

6 Conclusions

As the staining and intrinsic resolution of 2-DE continue to improve, the computational techniques will evolve towards more accurate and automatic ways of image analysis and quantification. One of the key developments is in the establishment of statistical norms for different cells and tissues from multiple experiments. This will bring an improvement in the sensitivity and robustness of the results, permitting the quantification of subtle differential protein expressions. The approach is underpinned by the availability of fast and accurate pixel based gel image registration techniques, and parallel advances in other imaging modalities that involve morphometric and appearance modelling. The necessity of comparing a large number of multiple gel pairs will trigger a fundamental change in the image processing pipeline, with more emphasis being placed on the use of intensity distribution, rather than spot location, as the primary means of image registration. This will permit a better modelling of the gel formation process as well as systematic errors such as current leakage and regional expression inhomogeneities. Under this framework, it will become natural to incorporate the statistical appearance model of individual spots, thus minimizing propagation of errors from one procedure downstream to the next.

Although 2-DE may remain unrivalled in its capability in simultaneously separating thousands of proteins in future years, its real impact on proteomics can only be realised when the processing bottleneck is fully resolved. This not only calls for rapid throughput in data processing, but also the deployment of fully automatic approaches in identifying intrinsic trends in different gel samples.

AWD is supported by the Medical Research Council. AWD would also like to thank Paramate Horkaew for his advice and work in the watershed transform and thin plate splines.

Received December 13, 2002
Revised March 27, 2003
Accepted March 27, 2003

7 References

[1] Hartl, D. L., Jones, E. W., *Genetics: Analysis of Genes and Genomes*, Jones and Bartlett, Sudbury 2001.

[2] Wilkins, M. R., Sanchez, J. C., Williams, K. L., Hochstrasser, D. F., *Electrophoresis* 1996, *17*, 830–838.

[3] Hochstrasser, D. F., in: Wilkins, M. R., Williams, K. L., Appel, R. D., Hochstrasser, D. F. (Eds.), *Proteome Research: New Frontiers in Functional Genomics*, Springer-Verlag, Heidelberg 1997, pp. 187–220.

[4] Jungblut, P., Thiede, B., Zimny-Arndt, U., Muller, E. C. *et al.*, *Electrophoresis* 1996, *17*, 839–847.

[5] Rabilloud, T., *Proteomics* 2002, *2*, 3–10.

[6] Flory, M. R., Griffin, T. J., Martin, D., Aebersold, R., *Trends Biotechnol.* 2002, *20*, S23–S29.

[7] Nishihara, J. C., Champion, K. M., *Electrophoresis* 2002, *23*, 2203–2215.

[8] Pietrogrande, M. C., Marchetti, N., Dondi, F., Righetti, P. G., *Electrophoresis* 2002, *23*, 283–291.

[9] Westbrook, J. A., Yan, J. X., Wait, R., Welson, S. Y. *et al.*, *Electrophoresis* 2001, *22*, 2865–2871.

[10] Anderson, N. L., Anderson, N. G., *Mol. Cell. Proteomics* 2002, *1*, 845–867.

[11] Rodbard, D., Chrambach, A., *Proc. Natl. Acad. Sci. USA* 1970, *65*, 970–977.

[12] Guo, X.-H., Chen, S.-H., *J. Chem. Phys.* 1990, *149*, 129–139.

[13] Unlu, M., Morgan, M. E., Minden, J. S., *Electrophoresis* 1997, *18*, 2071–2077.

[14] Yan, J. X., Devenish, A. T., Wait, R., Stone, T. *et al.*, *Proteomics* 2002, *2*, 1682–1698.

[15] Alban, A., Davis, S. O., Bjorkesten, L., Andersson, C. *et al.*, *Proteomics* 2003, *3*, 36–44.

[16] Herbert, B. R., Sanchez, J.-C., Bini, L., in: Wilkins, M. R., Williams, K. L., Appel, R. D., Hochstrasser, D. F. (Eds.), *Proteome Research: New Frontiers in Functional Genomics*, Springer-Verlag, Heidelberg 1997, pp. 13–34.

[17] Ireland, W. P., Sulston, K. W., Summan, M., *Electrophoresis* 2002, *23*, 1652–1658.

[18] Miller, M. D., Jr., Acey, R. A., Lee, L. Y., Edwards, A. J., *Electrophoresis* 2001, *22*, 791–800.

[19] Miura, K., *Electrophoresis* 2001, *22*, 801–813.

[20] Pleissner, K. P., Oswald, H., Wegner, S., in: Pennington, S. R., Dunn, M. J. (Eds.), *Proteomics: From Protein Sequence to Function*, BIOS Scientific, Oxford 2001, pp. 131–150.

[21] Anderson, N. L., Taylor, J., Scandora, A. E., Coulter, B. P. *et al.*, *Clin. Chem.* 1981, *27*, 1807–1820.

[22] Lemkin, P. F., in: Endler, T., Hanash, S. (Eds.), *Proceedings of Two-Dimensional Electrophoresis*, VCH Press, Weinheim 1989, pp. 53–57.

[23] Lemkin, P. F., Lipkin, L. E., *Comput. Biomed. Res.*, 1981, *14*, 407–446.

[24] Lemkin, P. F., Lipkin, L. E., *Comput. Biomed. Res.* 1981, *14*, 355–380.

[25] Lemkin, P. F., Lipkin, L. E., *Comput. Biomed. Res.* 1981, *14*, 272–297.

[26] Tarroux, P., Vincens, P., Meyer, J. A., in: Endler, T., Hanash, S. (Eds.), *Proceedings of Two-Dimensional Electrophoresis*, VCH Press, Weinheim 1989, pp. 68–71.

[27] Vincens, P., Tarroux, P., *Electrophoresis* 1987, *8*, 100–107.

[28] Garrels, J. I., *J. Biol. Chem.* 1989, *264*, 5269–5282.

[29] Kuick, R. D., Skolnick, M. M., Hanash, S. M., Neel, J. V., *Electrophoresis* 1991, *12*, 736–746.

[30] Rowlands, D. G., Flook, A., Payne, P. I., van Hoff, A. *et al.*, *Electrophoresis* 1988, *9*, 820–830.

[31] Olson, A. D., Miller, M. J., *Anal. Biochem.* 1988, *169*, 49–70.

[32] Raman, B., Cheung, A., Marten, M. R., *Electrophoresis* 2002, *23*, 2194–2202.

[33] Appel, R. D., Palagi, P. M., Walther, D., Vargas, J. R. *et al.*, *Electrophoresis* 1997, *18*, 2724–2734.

[34] Appel, R. D., Vargas, J. R., Palagi, P. M., Walther, D. *et al.*, *Electrophoresis* 1997, *18*, 2735–2748.

[35] Takahashi, K., Nakazawa, M., *Genome Inform.* 2001, *12*, 212–221.

[36] Takahashi, K., Nakazawa, M., Watanabe, Y., Konagaya, A., *Genome Inform.* 1998, *9*, 161–172.

[37] Takahashi, K., Nakazawa, M., Watanabe, Y., Konagaya, A., *4th Ann. Int. Conf. Computational Molecular Biology, Tokyo*, ACM, New York 2000, Poster 87.

[38] Baker, M., Busse, H., Vogt, M., *Medical Imaging 2000: Image Processing, San Diego*, SPIE, Bellingham 2000, *3979*, 426–436.

[39] Lemkin, P. F., *Electrophoresis* 1997, *18*, 461–470.

[40] Lemkin, P. F., Myrick, J. E., Lakshmanan, Y., Shue, M. J. *et al.*, *Electrophoresis* 1999, *20*, 3492–3507.

[41] Smilansky, Z., *Electrophoresis* 2001, *22*, 1616–1626.

[42] Miller, M. J., Merril, C., *Appl. Theor. Electrophor.* 1989, *1*, 127–135.

[43] Mahon, P., Dupree, P., *Electrophoresis* 2001, *22*, 2075–2085.

[44] Rogers, M., Graham, J., Tonge, R. P., *Proteomics* 2003, *3* (6), 879–886.

[45] Rogers, M., Graham, J., Tonge, R. P., *Proteomics* 2003, *3* (6), 887–896.

[46] Gustafsson, J. S., Blomberg, A., Rudemo, M., *Electrophoresis* 2002, *23*, 1731–1744.

[47] Tyson, J. J., Haralick, R. H., *Electrophoresis* 1986, *7*, 107–113.

[48] Sternberg, S. R., *Comput. Vis. Graph. Image Process.* 1986, *25*, 333–355.

[49] Skolnick, N. M., *Comput. Vis. Graph. Image Process.* 1986, *35*, 306–322.

[50] Wu, Y., Lemkin, P. F., Upton, K., *Electrophoresis* 1993, *14*, 1351–1356.

[51] Garrels, J. I., *J. Biol. Chem.* 1979, *254*, 7961–7977.

[52] Prehm, J., Jungblut, P., Klose, J., *Electrophoresis* 1987, *8*, 562–572.

[53] Conradsen, K., Pedersen, J., *Biometrics* 1992, *48*, 1273–1287.

[54] Horgan, G. W., Glasbey, C. A., *Electrophoresis* 1995, *16*, 298–305.

[55] Roerdink, J. B. T. M., Meijster, A., *Ann. Soc. Math. Pol. Ser. IV Fundam. Inf.* 2001, *41*, 187–228.

[56] Vincent, L., Soille, P., *IEEE Trans. Pattern Anal. Mach. Intell.* 1991, *13*, 583–598.

[57] Beucher, S., Lantuèjoul, C., *International Workshop on Image Processing, Rennes, France*, 1979, 2.1–2.12, http://cmm.ensmp.fr/~beucher/publi/watershed.pdf.

[58] Beucher, S., *Invited lecture, Société Italienne de Microscopie Electronique*, Taormina, Sicily, 1992, http://cmm.ensmp.fr/~beucher/publi/taorm.pdf. 1992.

[59] Pleissner, K. P., Hoffmann, F., Kriegel, K., Wenk, C. *et al.*, *Electrophoresis* 1999, *20*, 755–765.

[60] Li, S. Z., *Markov Random Field Modelling in Computer Vision*, Springer-Verlag, Heidelberg 1995.

[61] Besag, J., *J. R. Stat. Soc. Ser. B Stat. Methodol.* 1986, *48*, 269–302.

[62] Bettens, E., Scheunders, P., Van Dyck, D., Moens, L. *et al.*, *Electrophoresis* 1997, *18*, 792–798.

[63] Joliffe, I., *Principle Component Analysis*, Springer-Verlag, Heidelberg 1986.

[64] Lemkin, P. F., Myrick, J. E., Upton, K. M., *Appl. Theor. Electrophor.* 1993, *3*, 163–172.

[65] Efrat, A., Hoffmann, F., Kriegel, K., Schultz, C. *et al.*, *J. Comput. Biol.* 2002, *9*, 299–315.

[66] Dutt, M. J., Lee, K. H., *Electrophoresis* 2001, *22*, 1627–1632.

[67] Brown, L. G., *ACM Comput. Surv.* 1992, *24*, 325–376.

[68] Maintz, J. B., Viergever, M. A., *Med. Image Anal.* 1998, *2*, 1–36.

[69] Penney, G. P., Weese, J., Little, J. A., Desmedt, P. *et al.*, *Lect. Notes Comput. Sci.* 1998, *1496*, 1153–1161.

[70] Glasbey, C. A., Mardia, K. V., *J. Appl. Stat.* 1998, *25*, 155–171.

[71] Panek, J., Vohradsky, J., *Electrophoresis* 1999, *20*, 3483–3491.

[72] Pedersen, L. B. E., *Proc. 12th Scandinavian Conf. Image Analysis*, NOBIM, Tromso 2001, 118–125.

[73] Horgan, G., Creasey, A., Fenton, B., *Electrophoresis* 1992, *13*, 871–875.

[74] Salmi, J., Aittokallio, T., Westerholm, J., Griese, M. *et al.*, *Electrophoresis* 2002, *2*, 1504–1515.

[75] Akutsu, T., Kanaya, K., Ohyama, A., Fujiyama, A., *Lect. Notes Comput. Sci.* 1999, *1645*, 212–222.

[76] Alt, H., Guibas, J., in: Sack, J.-R., Urrutia, J. (Eds.), *Handbook of Computational Geometry*, B.V. North-Holland, Amsterdam 1999, pp. 121–153.

[77] Veltkamp, R. C., *Proc. 3rd Int. Conf. Shape Modeling and Applications*, IEEE, New York 2001, pp. 188–199.

[78] Pedersen, L., *PhD Thesis*, Technical University of Denmark, 2002.

[79] Toussaint, G. T., *Pattern Recognit.* 1980, *12*, 261–268.

[80] Wolfson, H. J., Rigoutsos, I., *IEEE Comput. Sci. Eng.* 1997, *4*, 10–21.

[81] Hoffmann, F., Kriegel, K., Wenk, C., *Discrete Appl. Math.* 1999, *93*, 75–88.

[82] Hoffmann, F., Kriegel, K., Wenk, C., *Proc. 14th Ann. Symp. Computational Geometry*, ACM, New York 1998, *14,* 231–239.

[83] Kriegel, K., Seefeldt, I., Hoffmann, F., Schultz, C. *et al.*, *Electrophoresis* 2000, *21*, 2637–2640.

[84] Gold, S., Rangarajan, A., Lu, C.-P., Pappu, S. *et al.*, *Pattern Recognit.* 1998, *31*, 1019–1031.

[85] Veeser, S., Dunn, M. J., Yang, G. Z., *Proteomics* 2001, *1*, 856–870.

[86] Beauchemin, S. S., Barron, J. L., *ACM Comput. Surv.* 1995, *27*, 433–467.

[87] Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P., *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, Cambridge 1997.

[88] Thevenaz, P., Unser, M., *IEEE Trans. Image Process.* 2000, *9*, 2083–2099.

[89] Velthuizen, R. P., Heine, J. J., Cantor, A. B., Lin, H. *et al.*, *Med. Phys.* 1998, *25*, 1655–1666.

[90] Likar, B., Maintz, J. B., Viergever, M. A., Pernus, F., *J. Microsc.* 2000, *197*, 285–295.

[91] Brinkmann, B. H., Manduca, A., Robb, R. A., *IEEE Trans. Med. Imaging* 1998, *17*, 161–171.

[92] Likar, B., Viergever, M. A., Pernus, F., *IEEE Trans. Med. Imaging* 2001, *20*, 1398–1410.

[93] Lai, S. H., Fang, M., *Med. Image Anal.* 1999, *3*, 409–424.

[94] Yang, G. Z., Myerson, S., Chabat, F., Pennell, D. J. *et al.*, *MAGMA* 2002, *14*, 39–44.

[95] Wells, W. M., III, Grimson, W. E. L., Kikinis, R., Jolesz, F. A., *IEEE Trans. Med. Imaging* 1996, *15*, 429–442.

[96] Bilmes, J.,Technical Report, *ICSI-TR-97-021*, International Computer Science Institute University of Berkeley 1997.

[97] Schmid, H. R., Schmitter, D., Blum, P., Miller, M. *et al.*, *Electrophoresis* 1995, *16*, 1961–1968.

[98] Pun, T., Hochstrasser, D. F., Appel, R. D., Funk, M. *et al.*, *Appl. Theor. Electrophor.* 1988, *1*, 3–9.

[99] Vohradský, J., *Electrophoresis* 1997, *18*, 2749–2754.

[100] Takahashi, K., Nakazawa, M., Watanabe, Y., Konagaya, A., *Genome Inform.* 1999, *10*, 121–132.

[101] Lemkin, P. F., Lester, E. P., *Electrophoresis* 1989, *10*, 122–140.

[102] Devore, J. L., *Probability and Statistics for Engineering and the Sciences*, Duxbury, Pacific Grove 2000.

[103] Picard, P., Bourgoin-Greneche, M., Zivy, M., *Electrophoresis* 1997, *18*, 174–181.

[104] Taylor, J., Giometti, C. S., *Electrophoresis* 1992, *13*, 162–168.

[105] Kettman, J. R., Robinson, R. A., Kuhn, L., Lefkovits, I., *Electrophoresis* 1991, *12*, 554–569.

[106] Kovarova, H., Radzioch, D., Hajduch, M., Sirova, M. *et al.*, *Electrophoresis* 1998, *19*, 1325–1331.

[107] Pleissner, K. P., Regitz-Zagrosek, V., Krudewagen, B., Trenkner, J. *et al.*, *Electrophoresis* 1998, *19*, 2043–2050.

[108] Harris, R. A., Yang, A., Stein, R. C., Lucy, K. *et al.*, *Proteomics* 2002, *2*, 212–223.

[109] Ward Jr., J. H., *J. Am. Stat. Assoc.* 1963, *58*, 236–244.

[110] Jessen, F., Lametsch, R., Bendixen, E., Kjaersgard, I. V. *et al.*, *Proteomics* 2002, *2*, 32–35.

[111] Wold, H., in: Kotz, S., Johnson, N. L. (Eds.), *Encyclopedia of Statistical Sciences*, Wiley, New York 1985, pp. 581–591.

[112] Vohradský, J., Li, X. M., Thompson, C. J., *Electrophoresis* 1997, *18*, 1418–1428.

[113] Martel, A. L., Moody, A. R., Allder, S. J., Delay, G. S. *et al.*, *Med. Image Anal.* 2001, *5*, 29–39.

[114] Janier, M. F., Mazzadi, A. N., Lionnet, M., Frouin, F. *et al.*, *Acad. Radiol.* 2002, *9*, 26–39.

[115] Buyya, R., *High Performance Cluster Computing Volume 1: Architectures and Systems*, Prentice Hall PTR, New Jersey 1999.

[116] Foster, I., Kesselman, C., Tuecke, S., *Int. J. Supercomput. Appl.* 2001, *15*, 200–222.

[117] Frey, J., Tannenbaum, T., Foster, I., Livny, M. *et al.*, *Cluster Comput.* 2002, *5*, 237–246.

[118] Zhou, S., Wang, J., Zheng, X., Delisle, P., *Softw. Pract. Exper.* 1993, *23*, 1305–1336.

[119] Heymann, E., Senar, M. A., Luque, E., Livny, M., *Lect. Notes Comput. Sci.* 2000, *1971*, 214–227.

[120] Buyya, R., *High Performance Cluster Computing Volume 2: Programming and Applications*, Prentice Hall PTR, New Jersey 1999.

[121] Foster, I., *Designing and Building Parallel Programs: Concepts and Tools for Parallel Software Engineering*, Addison-Wesley, Redwood City 1995.

[122] You, J., Shen, H., *AeroSense: Visual Information Processing VIII, Florida*, SPIE, New York 1998, *3387*, 212–218.

[123] Bairoch, A., in: Wilkins, M. R., Williams, K. L., Appel, R. D., Hochstrasser, D. F. (Eds.), *Proteome Research: New Frontiers in Functional Genomics*, Springer-Verlag, Heidelberg 1997, pp. 93–148.

[124] Hoogland, C., Sanchez, J. C., Walther, D., Baujard, V. *et al.*, *Electrophoresis* 1999, *20*, 3568–3571.

[125] Peitsch, M. C., Wilkins, M. R., Tonella, L., Sanchez, J. C. *et al.*, *Electrophoresis* 1997, *18*, 498–501.

[126] Evans, G., Wheeler, C. H., Corbett, J. M., Dunn, M. J., *Electrophoresis* 1997, *18*, 471–479.

[127] Lemkin, P. F., *Electrophoresis* 1997, *18*, 2759–2773.

[128] Rebhan, M., Prilusky, J., *Electrophoresis* 1997, *18*, 2774–2780.

[129] Malmström, L., Malmström, J., Marko-Varga, G., Westergren-Thorsson, G., *J. Proteome Res.* 2002, *1*, 135–138.

[130] Pleissner, K. P., Sander, S., Oswald, H., Regitz-Zagrosek, V. *et al.*, *Electrophoresis* 1997, *18*, 480–483.

[131] Appel, R. D., Bairoch, A., Sanchez, J. C., Vargas, J. R. *et al.*, *Electrophoresis* 1996, *17*, 540–546.

[132] Hoogland, C., Baujard, V., Sanchez, J. C., Hochstrasser, D. F. *et al.*, *Electrophoresis* 2002, *18*, 2755–2758.