

A statistical model for unwarping of 1-D electrophoresis gels

C.A. Glasbey

Biomathematics and Statistics Scotland
King's Buildings, Edinburgh, EH9 3JZ, Scotland

L. Vali

Medical Microbiology, University of Edinburgh
Medical School, Edinburgh, EH8 9AG, Scotland

& J.S. Gustafsson

Department of Mathematical Statistics, Chalmers University of Technology
Gothenburg, Sweden

Abstract

A statistical model is proposed, which relates density profiles in 1-D electrophoresis gels, such as those produced by pulsed-field gel electrophoresis (PFGE), to databases of profiles of known genotypes. The warp in each gel lane is described by a trend that is linear in its parameters plus a first-order autoregressive process, and density differences are modelled by a mixture of two normal distributions. Maximum likelihood estimates are computed efficiently by a recursive algorithm that alternates between dynamic time warping to align individual lanes and generalised-least-squares regression to ensure that the warp is smooth between lanes. The method, illustrated using PFGE of *E. coli* O157 strains, automatically unwarps and classifies gel lanes, and facilitates manual identification of new genotypes.

Key words: Autoregressive process, Dynamic programming, Image warping, Mixture distribution, Pulsed-field gel electrophoresis.

1 Introduction

Gel electrophoresis is a key technology in genomics. For example, pulsed-field gel electrophoresis (PFGE), amplified fragment length polymorphisms (AFLP), and denaturing gradient gel electrophoresis (DGGE) are among the methods used to characterise genetic variation. All

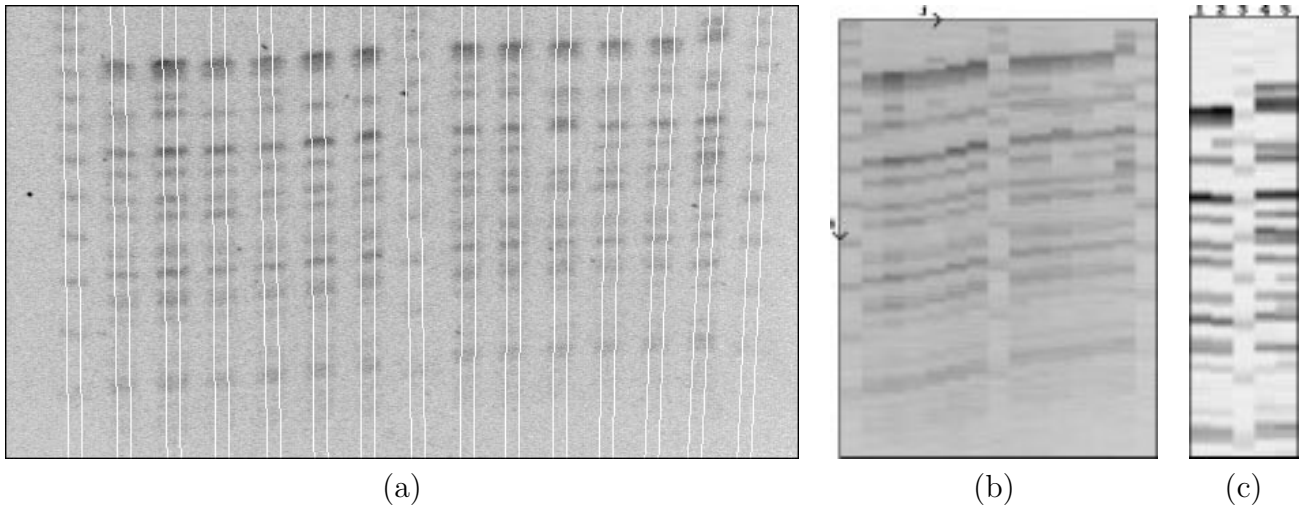


Figure 1: Pulsed-field gel electrophoresis of *E. coli* O157 strains: **(a)** original gel with manually-positioned parallel lines superimposed (lanes 1, 8, 15 are calibration lanes), **(b)** average intensities in each lane (Y^i), indexed by i and j , **(c)** database (μ) of lanes of four known genotypes and 3, a calibration lane.

these technologies generate similar data, so-called 1-D electrophoresis, as shown in Fig 1(a): each sample is represented by a vertical lane in the gel, with a density profile composed of a series of horizontal bands. Fig 1(a) is a PFGE of *Escherichia coli* O157:H7, a human pathogen that is carried and transmitted by cattle. Because human infection has been particularly prevalent in Scotland over the past decade, a study has been conducted of genotypic variation among *E. coli* O157 strains in Scottish cattle, to determine if they develop resistance to antibiotics [1]. Isolates obtained from naturally infected animals were typed by PFGE with XbaI restriction endonuclease enzyme [2, 3].

To identify bands in images such as Fig 1(a), and relate them to databases of profiles of known genotypes, some automation is possible [4], but typically, much human intervention is needed. This work is time consuming and results inevitably suffer from some subjectivity, so full automation is desirable. In §2, a statistical model is proposed for the density profiles on a single gel, and parameters are estimated by maximum likelihood using a fast algorithm. Then, in §3, the method is applied to three gels, giving automatic classification of strains and aiding the identification of new genotypes to augment the database. Finally, in §4, the method is discussed.

2 Materials and methods

2.1 Model formulation

Our first step is to reduce the size of gel datasets such as shown in Fig 1(a), and improve the signal-to-noise ratio. We compute lane profile intensities by cross-lane averaging between manually-positioned parallel lines, as shown in Fig 1(a) (though it would not be difficult to automate the positioning of these lines). In this example, the lanes were 11 pixels wide, and for simplicity, averages were computed along rows, although it would also be possible to average along transects perpendicular to the lanes. The result is an array, Y' , of J columns, each of length I , as illustrated in Fig 1(b), for which $J = 15$ and $I = 330$. We denote an individual element by Y'_{ij} .

Fig 2(a) shows Y'_i plotted against i for two lanes from different gels, where, for notational simplicity we will omit the j subscript. We see that the background intensity varies with i and differs between gels, and that band intensities are also different between gels. We correct for background trend t_i and scale s :

$$Y_i = \frac{Y'_i - \hat{t}_i}{\hat{s}}, \quad \text{where } \hat{t}_i = \max_{|p-i| \leq r} \left\{ \min_{|q-p| \leq r} Y'_q \right\}, \quad \hat{s} = \sqrt{\frac{1}{I} \sum_{i=1}^I (Y'_i - \hat{t}_i)^2},$$

and \hat{t} is a morphological closing of width $2r$, chosen to be greater than the width of individual bands (see, for example, Glasbey and Horgan [5]). Lanes, μ , in the database have been similarly standardised. Fig 2(b) shows Y and μ plotted against i . We can see that the transformation has been effective in correcting for differences in background trend and scale.

A second feature, evident in Fig 2(b), is that the i -axis needs to be stretched/warped to align bands in gel and database lanes. We propose

$$Y_i = \mu_{f_i} + \epsilon_i,$$

where f is the monotonic warp to align Y and μ . For a review of image warping methods, see Glasbey and Mardia [6, 7]. The least squares solution can be obtained by dynamic time warping, further details for which will be given in §2.2. Fig 2(c) shows the database lane after warping, from which the alignment with the gel lane can be seen to be very good, and Fig 2(e) shows warps of two lanes in one gel.

The model we propose is as follows:

$$Y_{ij} = \mu_{f_{ij}, l_j} + \epsilon_{ij}, \tag{1}$$

Here, l_j is the column in the database that matches column j in the gel, f_{ij} is the warp to align these two columns, and ϵ_{ij} is the sampling noise in the gel. We assume that the ϵ are independent and identically distributed with zero mean.

If the gel column is represented in the database, then we expect ϵ to be normally distributed with a small variance. However, for a new genotype whose band pattern differs from a closely

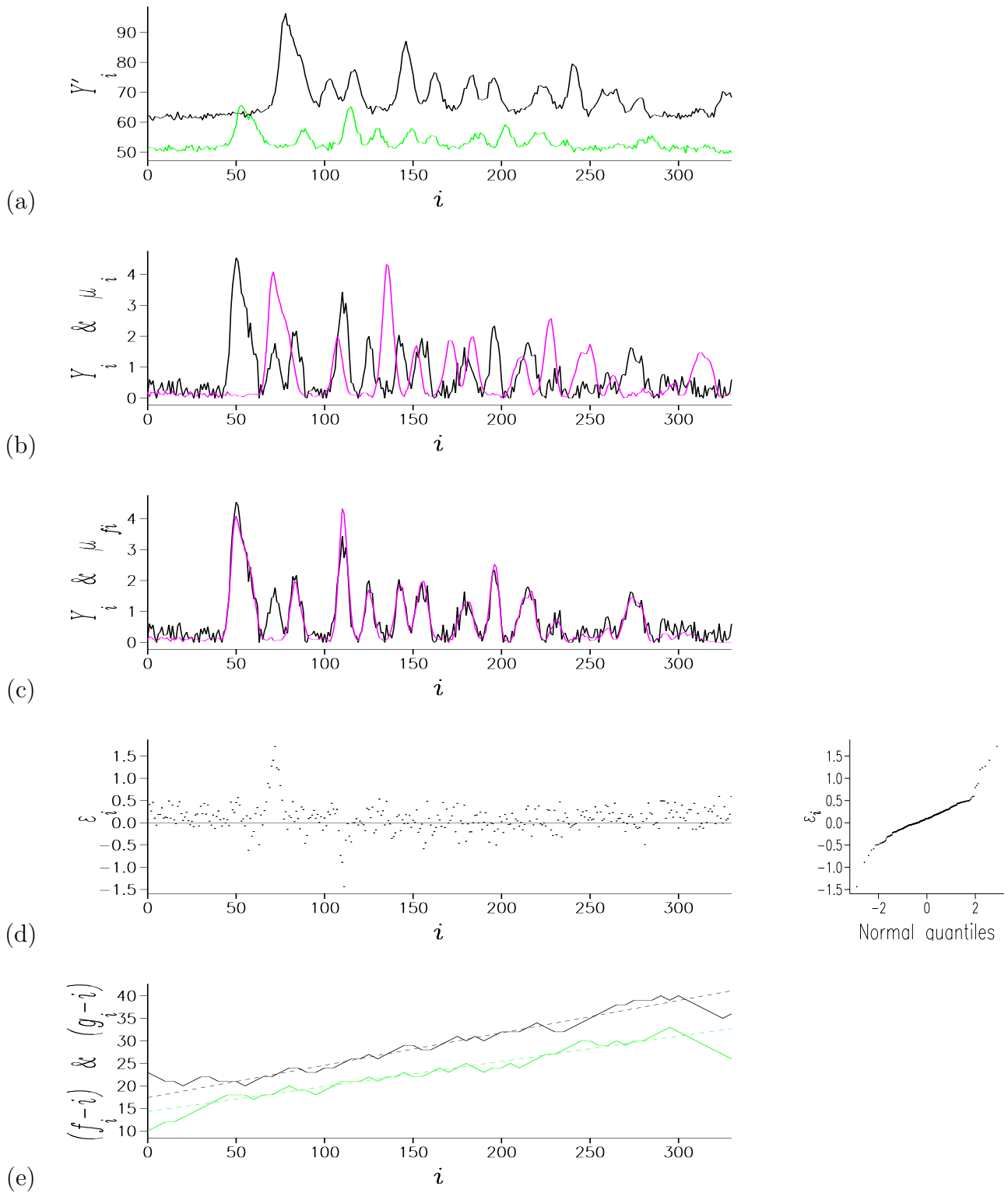


Figure 2: Exploratory plots against row index of pairs of lanes: **(a)** original intensities of two lanes from different gels (Y'), **(b)** intensities of one lane after trend and scale correction (Y) and one database lane (μ), **(c)** one lane together with warped database lane (μ_f), **(d)** difference between gel and warped database lanes (ϵ), and its normal probability plot, **(e)** warps (f) of two lanes in one gel, and linear approximations (g).

related genotype in the database only at a limited number of bands, ϵ will take larger values, as shown in Fig 2(d). Outliers can also result from image contamination, such as the dark specks visible in Fig 1(a). Therefore, we assume a mixture of two normal distributions

$$\epsilon \sim \begin{cases} N(0, \sigma_1^2) & \text{with probability } p \\ N(0, \sigma_2^2) & \text{otherwise} \end{cases}$$

which has probability density function

$$\mathcal{P}(\epsilon) = \frac{p}{\sqrt{2\pi\sigma_1^2}} e^{-\epsilon^2/2\sigma_1^2} + \frac{(1-p)}{\sqrt{2\pi\sigma_2^2}} e^{-\epsilon^2/2\sigma_2^2}.$$

Fig 2(e) also shows f approximated by a linear trend in each of two lanes. The smooth components are very similar in the two lanes, whereas fluctuations about these trends are correlated down each lane but appear to be uncorrelated between lanes. This leads us to propose that we decompose f into a component that is smooth across the whole gel and a component that is stochastic and uncorrelated between columns:

$$f_{ij} = g_{ij} + \eta_{ij}, \quad (2)$$

where g is linear in K parameters β , so $g_{ij} = \sum_k X_{ijk}\beta_k$ for some design matrix X , and η is an independent first-order autoregressive process in each column

$$\eta_{ij} \sim N(\phi\eta_{i-1,j}, \tau^2).$$

In a Bayesian formulation, we could view the model for f as a prior, as in [8]. Note, although in Fig 2(e) we show g as a straight line, our model formulation is far more general than that: we simply require g to be linear in its parameters, which encompasses polynomials, kernel and spline regression models (see, for example, Hastie and Tibshirani [9]). So, non-linear distortions are allowed.

2.2 Model fitting

The log-likelihood for Y given by (1) and (2), conditional on data in row $i = 0$, is:

$$\mathcal{L} = \sum_{j=1}^J \sum_{i=1}^I \log \mathcal{P}(\epsilon_{ij}) - \frac{IJ}{2} \log(2\pi\tau^2) - \frac{1}{2\tau^2} \sum_{j=1}^J \sum_{i=1}^I (\eta_{ij} - \phi\eta_{i-1,j})^2. \quad (3)$$

Alternatively, in a Bayesian formulation \mathcal{L} is the posterior log-density. In either case, (3) can be maximised efficiently using a recursive algorithm, to produce either the maximum likelihood estimator or the maximum *a posteriori* (MAP) estimator. The steps are as follows:

1) Estimate f and l with all other parameters held fixed:

$$\hat{f}_j, \hat{l}_j = \arg \max_{f_j, l_j} \sum_i \left[\log \mathcal{P}(Y_{ij} - \mu_{f_j, l_j}) - \frac{1}{2\tau^2} \{ (f_{ij} - g_{ij}) - \phi(f_{i-1,j} - g_{i-1,j}) \}^2 \right] \quad \text{for } j = 1, \dots, J.$$

This satisfies the *Principle of Optimality* [10] because it is a sum of separate costs, so its global solution can be found by dynamic programming. Algorithmic details are given in [8]. In this context it is called *Dynamic Time Warping*, and has been applied to speech processing [11] and handwriting analysis [12]. The algorithm also has similarities with that of Viterbi for fitting Hidden Markov Models [13]. Maximum likelihood estimation of l acts as an automatic classifier of each gel lane.

2) Estimate σ_1 , σ_2 and p , from $\epsilon_{ij} = Y_{ij} - \mu_{f_{ij}, l_j}$, by maximum likelihood using the EM-algorithm. We repeatedly solve the following until convergence:

$$\hat{\sigma}_1^2 = \frac{\sum_j \sum_i p_{ij} \epsilon_{ij}^2}{\sum_j \sum_i p_{ij}}, \quad \hat{\sigma}_2^2 = \frac{\sum_j \sum_i (1 - p_{ij}) \epsilon_{ij}^2}{\sum_j \sum_i (1 - p_{ij})}, \quad \hat{p} = \frac{1}{IJ} \sum_j \sum_i p_{ij},$$

where

$$p_{ij} = \frac{1}{\mathcal{P}(\epsilon_{ij})} \frac{p}{\sqrt{2\pi\sigma_1^2}} e^{-\epsilon_{ij}^2/2\sigma_1^2} \quad \text{for } j = 1, \dots, J, \quad i = 1, \dots, I.$$

3) Use generalised-least-squares regression to estimate τ , ϕ and β by maximising (3), with all other parameters held fixed, by repeatedly solving the following until convergence:

$$\sum_{k'=1}^K \sum_j \sum_i (X_{ijk} - \phi X_{i-1,jk})(X_{ijk'} - \phi X_{i-1,jk'}) \hat{\beta}_{k'} = \sum_j \sum_i (X_{ijk} - \phi X_{i-1,jk})(f_{ij} - \phi f_{i-1,j}),$$

$$\hat{\phi} = \frac{\sum_j \sum_i (f_{ij} - g_{ij})(f_{i-1,j} - g_{i-1,j})}{\sum_j \sum_i (f_{i-1,j} - g_{i-1,j})^2},$$

$$\hat{\tau}^2 = \frac{1}{IJ} \sum_j \sum_i \{(f_{ij} - g_{ij}) - \phi(f_{i-1,j} - g_{i-1,j})\}^2.$$

4) Repeat steps (1)–(3) until convergence.

Before starting the algorithm, we initialise parameter values in an *ad hoc* way:

0) Dynamic time warping of lanes with database, but simply using a sum of squares cost function

$$\hat{f}_j, \hat{l}_j = \arg \max_{f_j, l_j} \sum_i (Y_{ij} - \mu_{f_{ij}, l_j})^2 \quad \text{for } j = 1, \dots, J,$$

followed by robust regression

$$\hat{\beta} = \arg \max_{\beta} \sum_j \sum_i w_{ij} (f_{ij} - g_{ij})^2,$$

with w chosen to downweight large residuals.

To speed-up computations in step (1) and ensure monotonicity, we approximate f_j by a piecewise linear function with steps of length Δ , restrict f to take integer values at the join points, and constrain $|f_{i+\Delta, j} - f_{ij} - \Delta| \leq 1$. We have also modified step (3) accordingly, so that ϕ

gel	$\hat{\sigma}_1$	$\hat{\sigma}_2$	\hat{p}	$\hat{\phi}$	$\hat{\tau}$	$\hat{\beta}$			
1	0.059	0.188	0.65	0.936	0.472	-6.63	0.788	0.978	0.0029
2	0.200	0.481	0.82	0.954	0.410	0.34	0.779	1.017	0.0060
3	0.166	0.486	0.90	0.944	0.217	6.92	0.626	1.001	-0.0005

Table 1: Estimated parameter values for three gels.

denotes the autocorrelation at lag Δ . However, to keep the notation relatively simple, we do not give details.

After convergence, we can use the contribution to \mathcal{L} from each lane as a measure of how well the model fitted that lane. Let \mathcal{L}_j denote the contribution from lane j :

$$\mathcal{L}_j = \sum_{i=1}^I \log \mathcal{P}(\epsilon_{ij}) - \frac{I}{2} \log(2\pi\tau^2) - \frac{1}{2\tau^2} \sum_{i=1}^I (\eta_{ij} - \phi\eta_{i-1,j})^2. \quad (4)$$

Relatively low values of \mathcal{L}_j are indicative of poor fit, possibly because the sample in lane j is of a genotype not currently in the database. However, rather than automate this decision-making process, we prefer to rely on human judgement at this stage in the analysis.

3 Results

We used the algorithm in §2.2 in order to relate data from three PFGE of *E. coli* O157 strains, the second of which is shown in Fig 1(b), to the database shown in Fig 1(c). We assumed a step length $\Delta = 5$ in the dynamic time warping, and a bilinear function for the smooth component of the warp, so $g_{ij} = \beta_0 + i\beta_1 + j\beta_2 + ij\beta_3$. In all cases, the algorithm converged within 10 iterations. Table 1 shows estimated parameter values, and Fig 3 shows the results of unwarping each gel, i.e. displaying $Y_{f_{ij}^{-1},j}$, where f^{-1} is an inverse function in the first index, defined such that $f_{f_{ij}^{-1},j}^{-1} \equiv i$. Values of \hat{l} are shown at the top of each column.

Fig 4 shows the results of Fig 3 after reordering by \hat{l} , but with class 5 omitted as it only occurred once. Visual inspection shows four unusual lanes, indicated by arrows, which are redisplayed as a ‘new class’. These were also the lanes with the lowest values of \mathcal{L}_j as given in (4). The fifth lane on the second gel also had a low value of \mathcal{L}_j , but as can be seen in Fig 1(a), there was a dark speck at the top of this lane which appears to be due to noise rather than a genuine band. We were able to identify these four arrowed lanes as belonging to two further genotypes, to be added to the database.

4 Discussion

In §2 we proposed a statistical model which relates density profiles in 1-D electrophoresis gels to databases of profiles of known genotypes. We modelled the warp in each gel lane by a trend

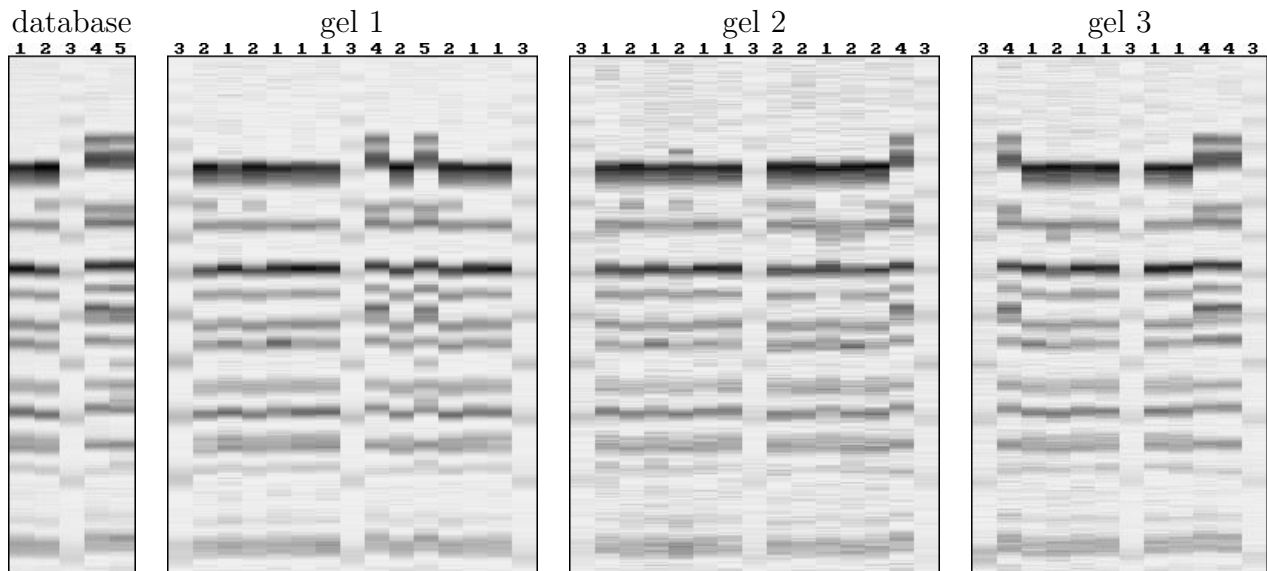


Figure 3: Database and three gels after unwarping.

that is linear in its parameters, common to all lanes, plus a first-order autoregressive process. Density differences were modelled by a mixture of two normal distributions. For this model formulation, maximum likelihood estimates can be computed efficiently by a recursive algorithm that alternates between dynamic time warping to align individual lanes and generalised-least-squares regression to ensure that the warp is smooth between lanes. The method was illustrated in §3. PFGE of *E. coli* O157 strains were automatically unwarped and gel lanes were classified, facilitating manual identification of new genotypes.

Dynamic programming is a fast, elegant method for finding the global solution to a class of 1-D optimisation problems, but unfortunately it does not generalise to higher dimensions. For the 1-D electrophoresis gels, as we only warp the row indices, we have a $1\frac{1}{2}$ -D warping problem, for which use can be made of dynamic programming to efficiently solve subproblems. However, 2-D electrophoresis is yet more challenging [14]

We note that although each the three steps in the algorithm in §2.2 finds a globally-optimal solution to a subproblem, the algorithm as a whole is not guaranteed to be globally-optimal. Therefore, choice of good initial values for parameters can be of critical importance. The use of stochastic optimisation algorithms, such as simulated annealing, may reduce the chances of finding a suboptimal solution, albeit at the cost of considerably greater computer time.

In further work we will test the algorithm on a larger dataset of PFGE gels and using gels from different technologies, and will automate the identification of new genotypes. The method is not restricted to profiles composed of bands and so has potential for many other applications.

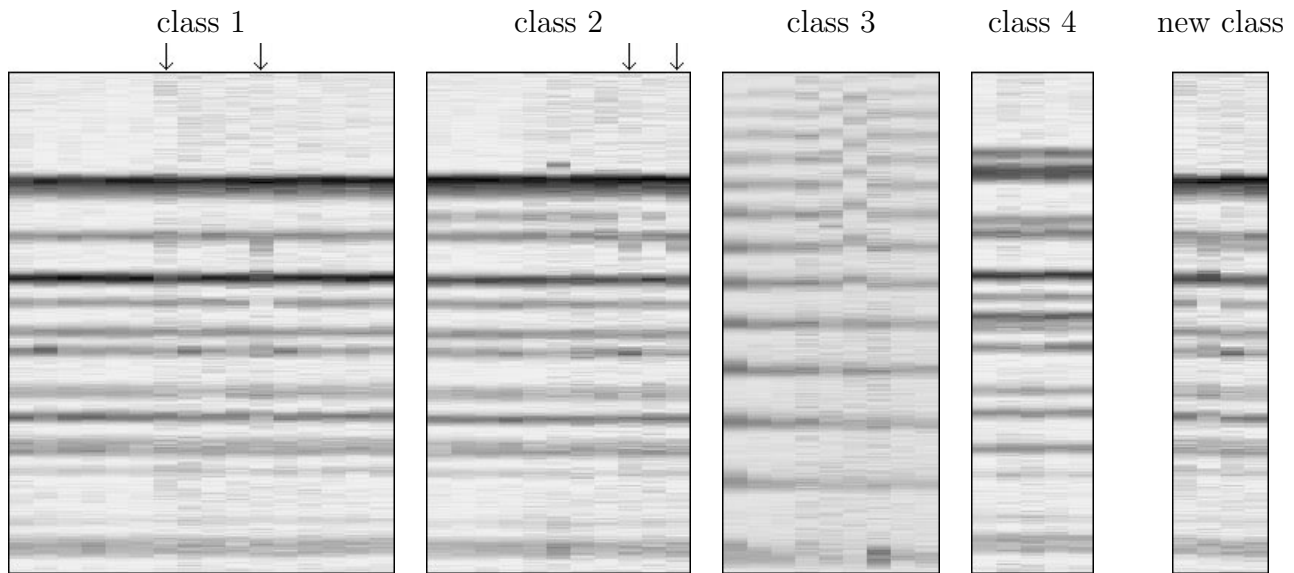


Figure 4: Columns in three unwarped gels, grouped by database class. Arrows indicate four outliers, also shown in ‘new class’.

Acknowledgements

CAG’s work was supported by funds from the Scottish Executive Environment and Rural Affairs Department. JG was supported by Chalmers Bioscience Programme. The data were part of the International Partnership Research Award in Veterinary Epidemiology (IPRAVE) funded by the Wellcome Trust.

References

- [1] L. Vali, K. A. Wisely, M. C. Pearce, E. J. Turner, H. I. Knight, A. W. Smith, and S. G. B. Amyes. High-level genotypic variation and antibiotic sensitivity among *Escherichia coli* O157 strains isolated from two Scottish beef cattle farms. *Applied and Environmental Microbiology*, 70:5947–5954, 2004.
- [2] T. J. Barrett, H. Lior, J. H. Green, R. Khakharia, J. G. Wells, B. P. Bell, K. D. Greene, J. Lewis, and P. M. Griffin. Laboratory investigation of a multistate food-borne outbreak of *Escherichia coli* O157:H7 by using pulsed-field gel electrophoresis and phage typing. *Journal of Clinical Microbiology*, 32:3013–3017, 1994.
- [3] R. K. Goutom. Rapid pulsed-field gel electrophoresis protocol for typing of *Escherichia coli* O157:H7 and other gram-negative organisms in 1 day. *Journal of Clinical Microbiology*, 35:2977–2980, 1997.
- [4] I. M. Skovgaard, K. Jensen, and I. Sondergaard. From image processing to classification: III. Matching patterns by shifting and stretching. *Electrophoresis*, 16:1385–1389, 1995.

- [5] C. A. Glasbey and G. W. Horgan. *Image Analysis for the Biological Sciences*. Wiley, Chichester, 1995.
- [6] C. A. Glasbey and K. V. Mardia. A review of image warping methods. *Journal of Applied Statistics*, 25:155–171, 1998.
- [7] C. A. Glasbey and K. V. Mardia. A penalised likelihood approach to image warping (with discussion). *Journal of the Royal Statistical Society, Series B*, 63:465–514, 2001.
- [8] C. A. Glasbey and M. J. Young. Maximum *a posteriori* estimation of image boundaries by dynamic programming. *Applied Statistics*, 51:209–221, 2002.
- [9] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman and Hall, London, 1990.
- [10] R. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, 1957.
- [11] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26:43–49, 1978.
- [12] D. J. Burr. Designing a handwriting reader. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5:554–559, 1983.
- [13] I. L. MacDonald and W. Zucchini. *Hidden Markov and Other Models for Discrete-Valued Time Series*. Chapman and Hall, London, 1997.
- [14] A. W. Dowsey, M. J. Dunn, and G.-Z. Yang. The role of bioinformatics in two-dimensional gel electrophoresis. *Proteomics*, 3:1567–1596, 2003.