

Warping of electrophoresis gels using generalisations of dynamic programming

Chris Glasbey

Biomathematics and Statistics Scotland

1 Introduction

Dynamic programming (DP) is a fast, elegant method for finding the global solution to a class of optimisation problems. For example, it can be used to align a single track in a 1-D electrophoresis gel, such as the pulsed-field gel electrophoresis (PFGE) shown in Fig 1(a), with another track. Gel tracks need to be aligned with tracks in a reference database, such as Fig 1(c), in order to genotype bacterial samples (Glasbey et al, 2005). However, it is not possible to use DP to align many PFGE tracks, for example to minimise

$$C(f, l) = \sum_i \sum_j [(Y_{ij} - \mu_{i+f_{ij}, l_j})^2 + \lambda_1(f_{ij} - f_{i-1, j})^2 + \lambda_2(f_{ij} - f_{i, j-1})^2]$$

with respect to a warping function (f) and track labels (l). Here Y is a two-dimensional array of gel pixel values, indexed by row i and track or column j , such as Fig 1(b), μ is a two-dimensional array of database pixel values, as in Fig 1(c), and C can be regarded as either a penalised log-likelihood or a Bayesian log-posterior density. The smoothness of the warp is determined by the magnitudes of λ_1 and λ_2 . To solve this optimisation problem we consider three generalisations of DP.

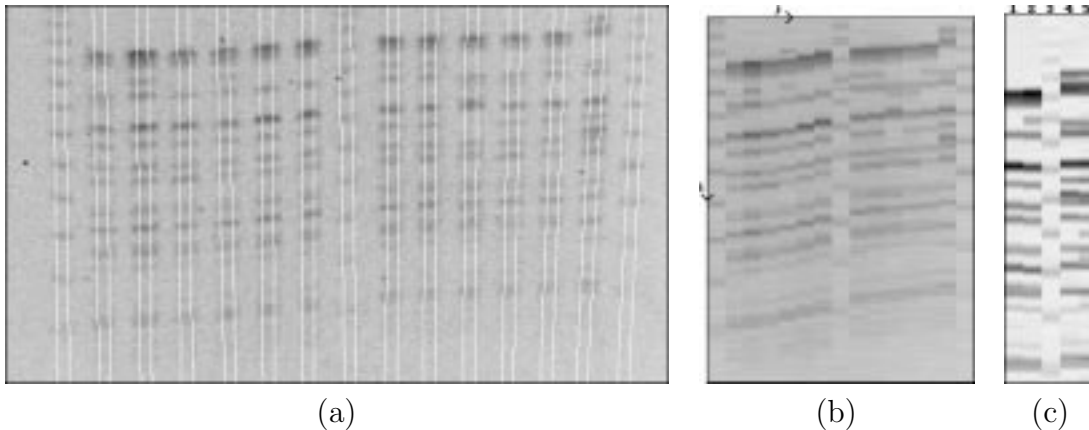


Figure 1: PFGE of *E. coli* O157 strains: **(a)** original gel with manually-positioned parallel lines superimposed, **(b)** intensities in each track (Y) after averaging between lines and normalising, **(c)** database (μ) of tracks of known genotype.

2 Generalisations of DP

The first generalisation is a greedy algorithm proposed by Leung et al (2004), termed iterated dynamic programming (IDP), where DP is used to recursively solve each of a

sequence of lower-dimensional problems in turn, to find a local optimum:

IDP

0) Initialise $f = 0$.

1) For each column j in turn, use DP to minimise $C(f, l)$ w.r.t. $f_{.j}$ and l_j , given current values of all other f 's and l 's, i.e. minimise

$$\sum_i \left[\left(Y_{ij} - \mu_{i+f_{ij}, l_j} \right)^2 + \lambda_1 (f_{ij} - f_{i-1, j})^2 + \lambda_2 \left\{ (f_{ij} - f_{i, j-1})^2 + (f_{ij} - f_{i, j+1})^2 \right\} \right]$$

where $(f_i - f_{i-1}) = 0.2, 0.4, \dots, 1.8$ for all i .

2) Repeat (1) until convergence.

A variant is to initialise f by applying DP independently to each column:

DP-IDP

0) For each column j in turn, use DP to minimise w.r.t. $f_{.j}$ and l_j

$$\sum_i \left[\left(Y_{ij} - \mu_{i+f_{ij}, l_j} \right)^2 + \lambda_1 (f_{ij} - f_{i-1, j})^2 \right].$$

A second algorithm is an empirical, stochastic optimiser, which is implemented by adding noise to IDP:

stoch-IDP

0) Initialise $f = 0$ and $T = 10^5$.

1) Simulate $e_{ijfl} \sim U[0, T]$ for all i, j, f, l . Then, for each column j in turn, use DP to minimise w.r.t. $f_{.j}$ and l_j

$$\sum_i \left[\left(Y_{ij} - \mu_{i+f_{ij}, l_j} \right)^2 + e_{ij, f_{ij}, l_j} + \lambda_1 (f_{ij} - f_{i-1, j})^2 + \lambda_2 \left\{ (f_{ij} - f_{i, j-1})^2 + (f_{ij} - f_{i, j+1})^2 \right\} \right]$$

2) Reduce $T \searrow \alpha T$, and repeat (1) until f unchanged, with α such that # iterations $\simeq 2^6, 2^7, \dots$

The final algorithm replaces DP by a more computationally intensive Forward-Backwards Gibbs Sampler (FB) (Scott, 2002), and uses a simulated annealing cooling schedule:

FB-SA

0) Initialise $f = 0$ and $T = 10^5$.

1) For each column j in turn, use FB to sample $f_{.j}$ and l_j with probability $\propto e^{-C(f, l)/T}$, given current values of all other f 's and l 's, i.e. sample from

$$\propto \exp \left[-\frac{1}{T} \sum_i \left(\left(Y_{ij} - \mu_{i+f_{ij}, l_j} \right)^2 + \lambda_1 (f_{ij} - f_{i-1, j})^2 + \lambda_2 \left\{ (f_{ij} - f_{i, j-1})^2 + (f_{ij} - f_{i, j+1})^2 \right\} \right) \right].$$

2) Reduce $T \searrow \alpha T$, and repeat (1) until f unchanged.

3 Results

In applying the algorithms to PFGE data such as Fig 1, we set $\lambda = (10, 1)$, which were the values identified by Gustafsson (2005) using trackwise cross-validation. Fig 2 shows the minimised values of C plotted against CPU time for multiple runs of the algorithms applied to the data in Fig 1(b) and to a second gel. We see that the deterministic algorithm DP-IDP performs best. Both stochastic algorithms (stoch-IDP, FB-SA) find the same optimal solution for gel 1, provided the cooling rate is not too rapid, but both appear to become trapped in local optima for gel 2. We also note that stoch-IDP outperforms FB-SA even though it lacks a theoretical justification. Finally, Fig 3 show the results of unwarping the two gels using the estimated values of f , and the estimated track labels \hat{l} .

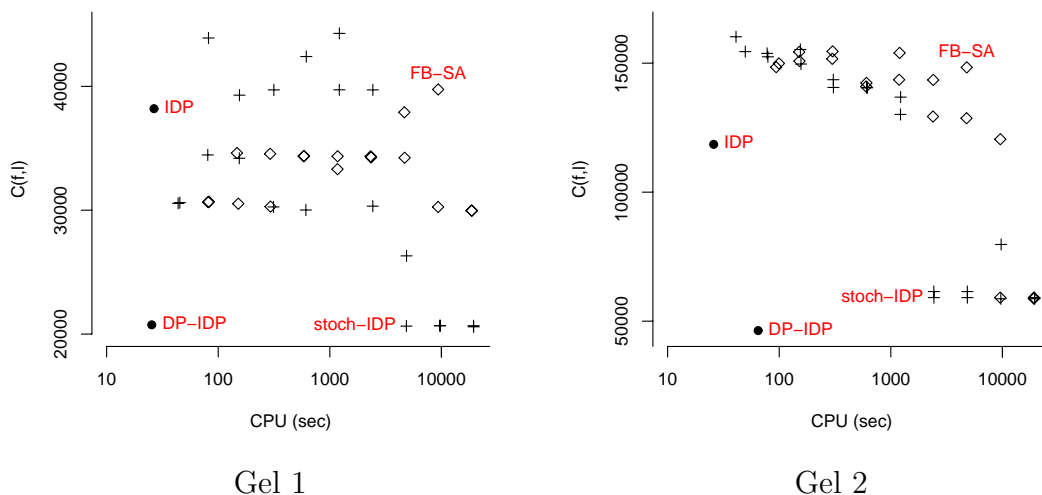


Figure 2: Minimised values of C plotted against CPU time for runs of algorithms on two gels.

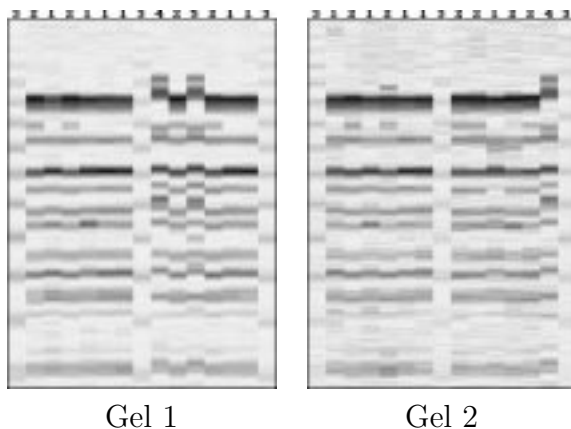


Figure 3: Unwarped gels, labelled with most similar tracks in database.

4 Discussion

In addition to 1-D warping, DP can also be used to find maximum a posteriori (MAP) estimators of boundaries, to automatically segment 2-D medical images into anatomical regions (Glasbey and Young, 2002). However, for many image problems, including 3-D segmentation and image restoration, as well as image warping, DP is not possible. We have also applied the generalisations of DP to restore synthetic aperture radar (SAR) remotely-sensed images and to segment 3-D X-ray computed tomographs. Further work is needed to evaluate the algorithms.

References

- Glasbey, C.A. and Young, M.J. (2002). Maximum a posteriori estimation of image boundaries by dynamic programming. *Applied Statistics*, **51**, 209-221.
- Glasbey, C.A., Vali, L. and Gustafsson, J.S. (2005). A statistical model for unwarping of 1-D electrophoresis gels. *Electrophoresis*, **26**, 4237-4242.
- Gustafsson, J. (2005). *Unwarping and Analysing Electrophoresis Gels*. PhD thesis, Chalmers University of Technology, Sweden.
- Leung, C., Appleton, B. and Sun, C. (2004). Fast stereo matching by Iterated Dynamic Programming and quadtree subregioning. *British Machine Vision Conference*, **1**, 97-106. Edited by A. Hoppe, S. Barman and T. Ellis.
- Scott, S.L. (2002). Bayesian methods for Hidden Markov Models: recursive computing in the 21st century. *Journal of the American Statistical Association*, **97**, 337-351.