A New Approach for Multi-Dimensional Signal Processing and Modelling for Signals From Gel Electrophoresis

Ebtihal H. G. Yousif, James R. Hopgood, John S. Thompson and Mike E. Davies
Institute for Digital Communications
School of Engineering
University of Edinburgh, Edinburgh, UK.
Emails: {e.yousif; james.hopgood; john.thompson; mike.davies}@ed.ac.uk

***Abstract*—In this paper, a multi-channel multi-dimensional approach is investigated for modeling of signals obtained from DNA gel electrophoresis. Related applications include DNA fingerprinting and crime scene investigations. In order to improve resolution and accuracy of modeling, a novel approach is employed based on using equidistant multi-capture data frames obtained over an extended span of time. The multidimensional signal is rescaled and aligned which improves resolution, then the signal is modeled as a surface that varies with both the time index and separation size. The overall approach is tested on a number of datasets. The simulation results show that the proposed approach can be used as a starting multi-dimensional time series model for raw signals obtained from gel electrophoresis.**

I. Introduction

Deoxyribonucleic acid (DNA) is a molecule that carries the genetic information and instructions in living beings. There are various tools and signal processing approaches that are used for analyzing biological molecules. One powerful tool in particular is Electrophoresis [1], which is applied for separating macromolecules based on their size, for purposes of identification, quantification or purification. The process of electrophoresis involves subjecting the DNA molecules to an electric field through a medium, e.g., a special gel, which makes the molecules move and separate based on their charge and size. The study of signals obtained from DNA gel electrophoresis are of significant importance, especially given that the majority of the existing literature focuses on the process of determining the sequence of chemical bases in a particular DNA molecule, rather than time series analysis of the captured signals from electrophoresis and/or statistical signal processing of those signals.

It is a well-known fact that inside a solution, a DNA molecule demonstrates the behavior and the characteristics of a Brownian particle [2]. In addition, the entire procedure of gel electrophoresis can be studied as a macrotransport process [3]. In fact, macrotransport processes are applicable to various cases and phenomena that arise in physiochemical systems. However, some of the involved parameters, or captured signals, may exhibit a stochastic nature and therefore can be studied using statistical signal processing approaches or time series analysis tools.

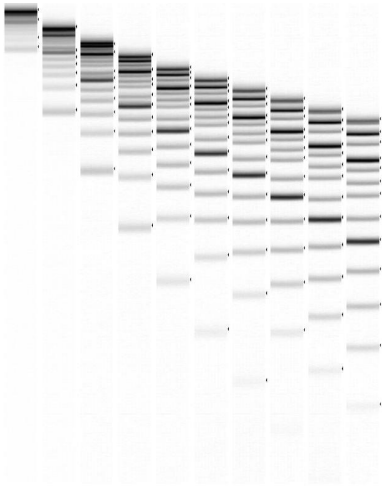


Fig. 1. Multiple Snapshots illustrating separation of DNA fragments

We would like to emphasize the point that in this paper, we aim to link statistical signal processing and analysis approaches to model the captured signals that result from gel electrophoresis. That is to say that we are not considering a genomic signal processing approach [4], i.e., we are not necessarily seeking the biological translation of captured data. Instead, we study the characteristics of the considered phenomena as a time series. Specifically, we aim to find a model that describes the resultant signal and its characteristics using a multidimensional approach. This is achieved by exploring the stochastic nature of signals captured from DNA gel electrophoresis, where the DNA molecules take a stochastic trajectory when separating during electrophoresis.

Basecalling is defined as the extraction or decoding of the DNA sequence from the processed time series. The available literature usually targets signal processing approaches for DNA sequencing and basecalling, but not the time series modeling of the obtained signal from electrophoresis. In [5], the analysis of the information content of linear biomolecules, such as DNA and proteins, is investigated via digital signal processing approaches. The study in [6] uses the wavelet

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

transform in DNA sequence analysis and on cellular neural networks in microarray image analysis, which can have a potentially large effect on the real-time realization of DNA analysis. The study also summarizes possible research approaches including a signal processing technique for genomic feature extraction and hybrid multidimensional approaches to process the dynamic genomic information in real time. Furthermore, in [7], the authors considered the computation of linear transforms of symbolic signals. The investigated techniques are illustrated by considering spectral and wavelet analyses of DNA sequences. Also, the study in [8] considered the recovery of signals in compressed DNA microarrays using sparse measurement matrices. Finally, in [9], a new algorithm for examining periodic patterns in DNA sequences is developed, using the short-time periodicity transform. Regarding mathematical modeling for DNA processing, some studies exist, e.g., [10]. Image processing-based approaches exist such as [11]–[13], but are mostly related to DNA microarray images.

In this paper we propose for the first time an approach for multi-channel multi-dimensional modeling of raw signals from DNA gel electrophoresis. The signal is passed through several processing stages that increases the resolution. First the peaks of the signal are detected, then the frames are scaled and realigned, and finally the multi-dimensional signal is fitted to a proposed theoretical model. The datasets used for testing our approach are based on a new novel multi-snapshot imaging approach [14], which provides higher resolution. Fig.1 illustrates an example of a multi-snapshot captured image, which demonstrates the progression of separation of the DNA fragments as time goes on.

The rest of this paper is organized as follows. Section II describes the preliminary background of the investigated problem. In section III, the signal modeling approach is explained. In section IV, simulation results are presented and finally section V concludes the paper.

II. Problem Formulation

In this section, preliminary information is provided and the investigated problem is introduced and formulated.

A. Mathematical Notations

The following notations will be used throughout the paper. Vectors and matrices will be denoted by lowercase and uppercase boldface characters, respectively. The notation $\hat{(\cdot)}$ indicates an estimated parameter, whereas $\bar{(\cdot)}$ indicates the statistical mean. Other mathematical operators include $(\cdot)^\top$ which is the transpose, $\|\cdot\|_p$ is the matrix p-norm, $\otimes$ is the Kronecker product, $\mathbf{1}_a$ is the $a \times 1$ ones vector and finally the notation $\mathcal{I}(\cdot)$ is the indicator function.

B. Problem Setup

Let N_D be the size of the DNA population that will be introduced at the start of the experiment, i.e., at $t = t_0$, where $t = 0, \ldots, N-1$. It is assumed that the stochastic trajectory of each DNA molecule can be described in terms of the cylindrical coordinates, where each point is described by the tuple $\mathbf{r} = \{r, \phi, z\}$, where

$$\mathbf{r} = \left\{ r, \phi, z \middle| 0 \le r \le r_{\max}, 0 \le \phi \le 2\pi, -\infty \le z \le \infty \right\}. \quad (1)$$

We are mainly interested in the average area concentration field of the DNA solute, which can be interpreted as a time series signal. Let $\mathcal{F}$ be the concentration of the DNA population in molecules per unit volume. Using the cylindrical coordinates, the initial concentration field at time t_0 is denoted by

$$\mathcal{F}_0\left(r_0, \phi_0, z_0\right) = \mathcal{F}\left(r_0, \phi_0, z_0, t = 0\right), \quad (2)$$

where the starting coordinate point is $\mathbf{r}_0 = \{r_0, \phi_0, z_0\}$ at time t_0. The initial population size is governed by

$$N_{\mathrm{D}} = \int_{-\infty}^{\infty} \int_{0}^{2\pi} \int_{0}^{r_{\max}} \mathcal{F}_0\left(r_0, \phi_0, z_0, t_0\right)\, r_0 \mathrm{d}r_0\, \mathrm{d}\phi_0\, \mathrm{d}z_0. \quad (3)$$

Furthermore, the concentration at time $t = 0$ is a function of the transition probability, i.e.,

$$\mathcal{F}\left(r, \phi, z, t\right) = \iiint \Pr\left(r, \phi, z, t \middle| r_0, \phi_0, z_0, t = 0\right) \times \mathcal{F}_0\left(r_0, \phi_0, z_0, t_0\right)\, r_0 \mathrm{d}r_0\, \mathrm{d}\phi_0\, \mathrm{d}z_0, \quad (4)$$

where $\Pr\left(r, \phi, z, t \middle| r_0, \phi_0, z_0, t_0\right)$ is the transition probability into a new state described by the coordinates $\mathbf{r} = \{r, \phi, z\}$ and time t and satisfying

$$\int_{-\infty}^{\infty} \int_{0}^{2\pi} \int_{0}^{r_{\max}} \Pr\Big(r, \phi, z, t \Big| r_0, \phi_0, z_0, t_0\Big) \times r \mathrm{d}r\, \mathrm{d}\phi\, \mathrm{d}z = \mathcal{I}(t \ge 0) \quad (5)$$

It ia assumed that the concentration $\mathcal{F}$ should satisfy the following conditions [3]:

$$\mathcal{F} \to 0, \qquad \text{at } |z| \to \infty, \quad (6a)$$

$$\mathcal{F} \to \mathcal{F}_0, \qquad \text{at } t = 0, \quad (6b)$$

$$\frac{\partial \mathcal{F}}{\partial r} = 0, \qquad \text{at } r = R, \quad (6c)$$

$$\mathcal{F}\left(\phi + 2\pi\right) = \mathcal{F}(\phi). \quad (6d)$$

The first condition stated by (6a), implies that the concentration field vanishes at large distances along the direction of movement. The second condition implies that the concentration is equivalent to the initial value at $t = t_0$. The condition in (6c) indicates that the rate of change of the concentration at the maxim value of the r axis is zero. Finally, the condition in (6d) states that the concentration is cyclic as a function of the ϕ axis. Furthermore, the concentration field should satisfy the convective-diffusion partial differential equation [3], which is written as

$$\frac{\partial \mathcal{C}}{\partial t} + v \frac{\partial \mathcal{C}}{\partial z} = \omega \left\{ \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial \mathcal{C}}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 \mathcal{C}}{\partial \phi^2} + \frac{\partial^2 \mathcal{C}}{\partial z^2} \right\}, \quad (7)$$

where v and ω denote the components of the mean velocity vector and the dispersion tensor in the direction of motion, i.e., along the z axis. Thus, the previous equation describes

what happens to the concentration $\mathcal{F}$ at long times. Solving the previous equation based on the initial conditions stated by equations (3) and (6) yields a *weighted* version of the traditional solution

$$\mathcal{F}(t,y) = \frac{1}{\sqrt{4\pi\omega t}} \exp\left[-\frac{(y-vt)^2}{4\omega t}\right]. \tag{8}$$

III. Signal Modeling

A. Organization of the Captured Signal

Let N_{F} denote the number of captured frames within the time window, and let M denote the number of separation indexes. Let us assume that the results from gel electrophoresis of a specific DNA fragment are organized into the matrix $\mathbf{Z} \in \mathbb{R}^{M \times N_{\mathrm{F}}}$, where

$$\mathbf{Z} = \Big\{ z_{m,t} | m = 1, \ldots, M, t = 0, \ldots, N_{\mathrm{F}} - 1 \Big\}, \tag{9}$$

where m and t represent the separation and time indexes respectively. The matrix $\mathbf{Z}$ can also be defined as a group of N_{F} objects in an M-dimensional space. Let us define $\mathbf{z}_t \in \mathbb{R}^{M \times 1}$, where $\mathbf{z}_t = \{z_{m,t}\}_{m=1}^{M}$, be the column vector of the obtained signal values at the t-th time instant. Hence, we can rewrite (9) on the form:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_0 & \mathbf{z}_1 & \ldots & \mathbf{z}_{N_{\mathrm{F}}-1} \end{bmatrix}. \tag{10}$$

Let us assume that the signal is passed through the following steps: (1) identification of peak values and their corresponding locations in the time-separation grid; (2) alignment of captured frames to a chosen frame; and (3) applying curve fitting using non-linear least squares to estimate a multidimensional model. Let us assume that the aligned matrix is described by the set of shift and scale gain parameters for each frame. Let us define the aligned matrix $\breve{\mathbf{Z}}$ as

$$\breve{\mathbf{Z}} = \mathbf{Z}\,\boldsymbol{\Xi} + \mathbf{1}_M \otimes \boldsymbol{\beta} \tag{11}$$

where $\boldsymbol{\Xi} \in \mathbb{R}^{N_{\mathrm{F}} \times N_{\mathrm{F}}}$ and $\boldsymbol{\beta} \in \mathbb{R}^{N_{\mathrm{F}} \times 1}$ denote a diagonal scaling matrix and the offset row vector respectively, which are defined by

$$\boldsymbol{\Xi} = \begin{bmatrix} \xi_1 & 0 & \ldots & 0 \\ 0 & \xi_2 & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & \xi_{N_{\mathrm{F}}} \end{bmatrix}, \tag{12}$$

and

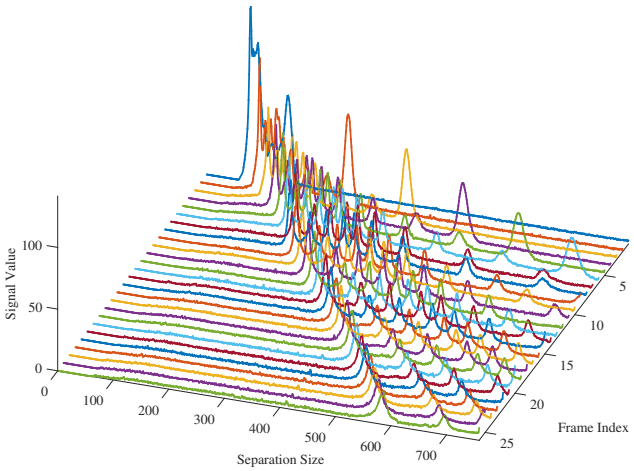
$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 & \beta_2 & \ldots & \beta_{N_{\mathrm{F}}}, \end{bmatrix} \tag{13}$$

respectively. Let $\boldsymbol{\rho}$ denote the set of peaks that is associated with the t-th frame $\mathbf{z}_t$, where $t = 0, \ldots, N_{\mathrm{F}} - 1$. Each list of peaks is defined as a set containing ordered pairs, such that the signal value can be characterized by the set of sets having the form

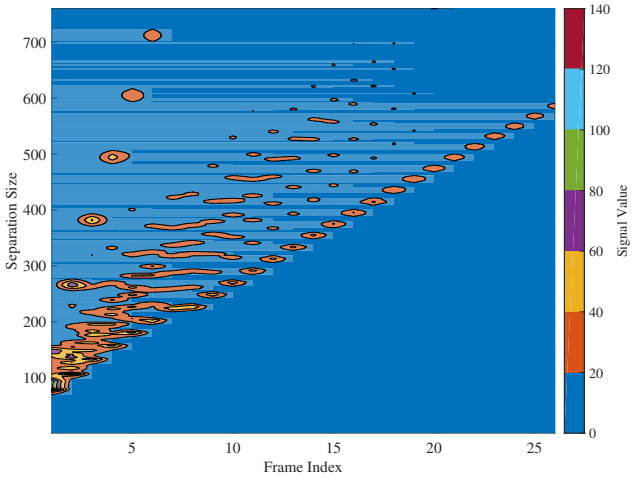
$$\boldsymbol{\mathcal{P}} = \Big\{ \{\rho_1\}, \ldots, \{\rho_{N_{\mathrm{F}}}\} \Big\} \tag{14}$$

where $\boldsymbol{\rho}_i$ denotes the i-th peaks list, which is defined as

$$\boldsymbol{\rho}_i = \Big\{ \big(i, m, \breve{z}_{(m,t=i)}\big)_k \,\big|k = 1, \ldots, K_i, \quad K_i < M,\ \breve{z} \in \big\{\breve{z}_{(m,t=i)}\big\}_{m=1}^{M}, \Big\},\ i = 1, \ldots, N_{\mathrm{F}}, \tag{15}$$



(a) Multichannel signal (raw signal)



(b) Contour view

Fig. 2. Sample datset obatined from DNA gel electrophoresis

where K_i is the number of peaks per frame. Henceforth, the entire alignment problem is to compress (or decompress) each frame, based on a chosen reference frame, and to find the values of scaling gain matrix $\boldsymbol{\Xi}$ and the shifting gain vector $\boldsymbol{\beta}$.

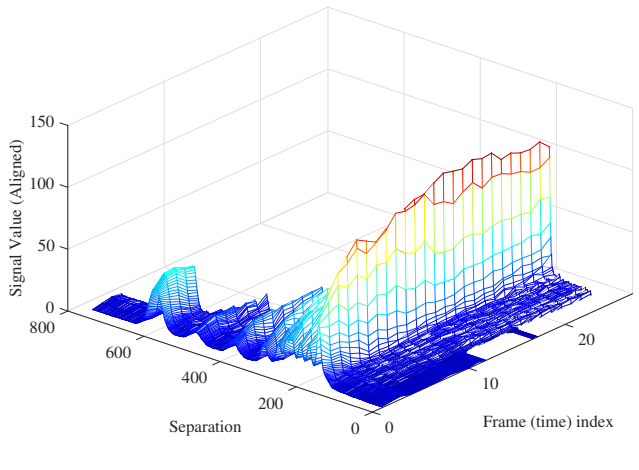
B. Multidimensional Model

In this part, we consider a theoretical model of the process $\{z_{m,t} | m = 1, \ldots, M; t = 0, \ldots, N_{\mathrm{F}} - 1\}$ as a two dimensional process denoted by

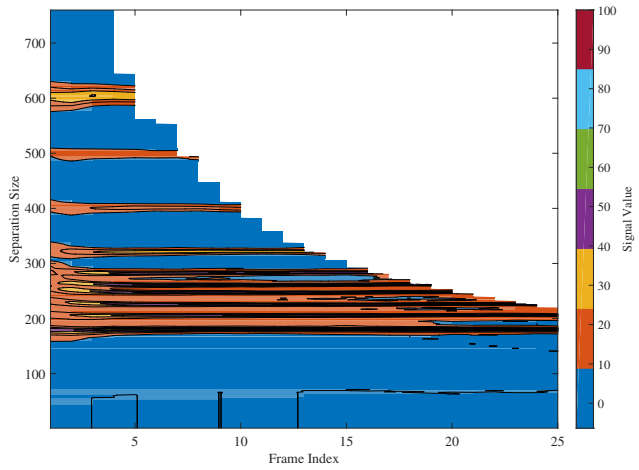
$$\widehat{z} = f(x,y) = \mathcal{F}(x = t, y) \tag{16}$$

Hence, we seek a theoretically estimated matrix $\hat{\mathbf{Z}}$ that is equivalent to the aligned data matrix $\breve{\mathbf{Z}}$. First, let us assume that each element of $\hat{\mathbf{Z}}$ is governed by the relationship:

$$\widehat{z} = \sum_{i=1}^{K_{\mathrm{avg}}} \frac{\hat{\alpha}_i(x)}{\sqrt{4\pi\hat{\omega}_i(x)}} \exp\left[-\frac{\Big(y - \hat{v}_i(x)\Big)^2}{4\hat{\omega}_i(x)}\right], \tag{17}$$

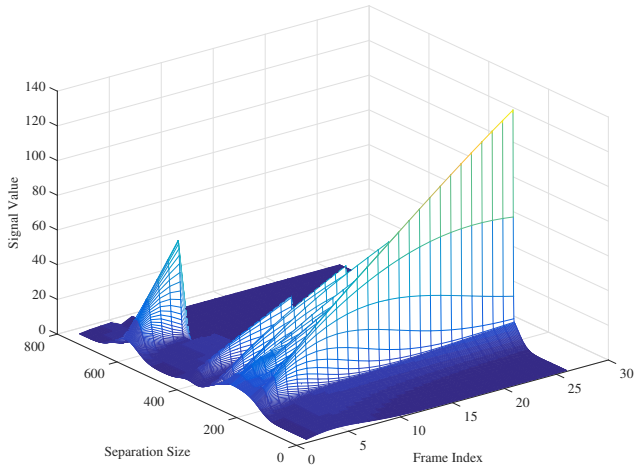


(a) Result of scaling and realignment

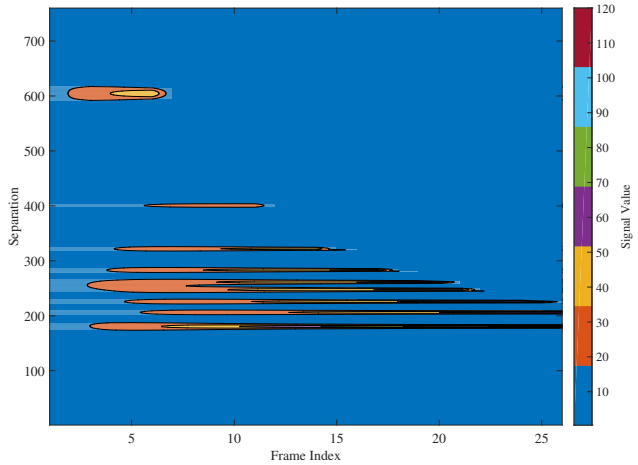


(b) Contour view showing impact of scaling and realignment

Fig. 3. Signal after scaling and realignment based on (11)



(a) Fitted signal



(b) Contour view of surface

Fig. 4. Synthetic Surface in (a)3D, (b) contour view. Fitting was perrfomed based on (18)

where in this case the parameters $\hat{\omega}$ and $\hat{v}$ are both functions of x, i.e., varying with time. The values of the estimated parameters can be obatined using a non-linear least squares formulation. This can be visualized as finding the optimal vector or parameters that satisfies

$$\begin{aligned} \tilde{\boldsymbol{\theta}} = \quad & \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left\| \check{\mathbf{Z}}(\boldsymbol{\theta}) - \hat{\mathbf{Z}}(\boldsymbol{\theta}) \right\|^2, \\ \text{subject to} \quad & \tilde{\boldsymbol{\theta}} \in \mathbb{R}_{+}^{3K_{avg}\times 1} \end{aligned} \tag{18}$$

where $\tilde{\boldsymbol{\theta}}$ is the optimal set of parameters which is defined as

$$\boldsymbol{\theta} = \operatorname{vec}(\boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{v}), \tag{19}$$

where $\boldsymbol{\alpha}$, $\boldsymbol{\omega}$ and $\boldsymbol{v}$ are all in $\mathbb{R}^{K_{avg}\times 1}$, and defined as

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 & \dots & \alpha_{K_{\text{avg}}} \end{bmatrix}^\top, \tag{20}$$

$$\boldsymbol{\omega} = \begin{bmatrix} \omega_1 & \dots & \omega_{K_{\text{avg}}} \end{bmatrix}^\top, \tag{21}$$

$$\boldsymbol{v} = \begin{bmatrix} v_1 & \dots & v_{K_{\text{avg}}} \end{bmatrix}^\top. \tag{22}$$

IV. Simulation Results and Discussion

In this section, examples from the simulation results are presented and discussed. In fact, the proposed approach is tested on several datasets. However, for brevity and due to limited space, only results for a single dataset are highlighted. The example dataset is depicted in Fig.2. The dataset was first obtained using a multi-capture imaging method, such that the signal is obtained as a surface that varies with both time index and separation size. The multiple time series captures are illustrated in Fig.2-(a) which shows the signal values variation across time and separation. In fact, using a multi-capture approach provides a higher resolution when compared with the conventional methods, namely the finish-line method and the single-snapshot imaging approach. On the other hand, Fig.2-(b) depicts a contour plot, which provides a top view that demonstrates the behavior of peaks as a function of both the time index and separation.

The results of the rescaling and realignment step are shown in Fig.3, where Fig.3-(a) illustrates the resultant surface and

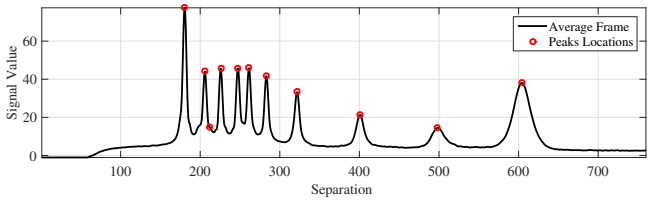


Fig. 5. Initial number of estimated average peak locations

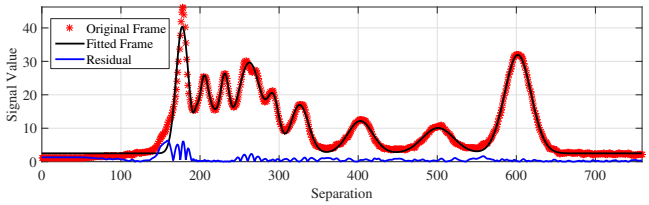


Fig. 6. Theoretical fitting of the starting frame based on the initial peaks locations

Fig.3-(b) provides a contour view. Both parts of the figure depicts the signal value as a function of both time index and separation size. In fact, rescaling and realignment of signal peaks provides higher resolution, but at an expense of missing information for specific coordinates of time and separation. The missing areas are indicated by the blank white region in figures (a) and (b) of Fig.3. The aligned matrix $\check{\mathbf{Z}}$ was obtained by finding the associated parameters defined in (11).

Finally, the fitting results are shown in Fig.4. Fig.4-(a) depicts the estimated surface obtained by solving the optimization problem of (18). Fig.4-(b) compares between the contour views of the original surface and the fitted surface. Future work to improve the accuracy of the procedure includes using advanced algorithms for baseline and background noise estimation.

For fitting, a list of average peak locations are calculated based on the results obtained from determining the multiset in (14). In the employed example dataset, the average number of peaks was found to be 11, with one false peak. Fig.5 illustrates the estimated average frame and the average peak locations. The set of initial peaks locations, was used to generate initial values for the optimization problem described by (18). Fig.6, depicts an example for the initial frame, showing the original signal as a function of separation, the fitted frame and the corresponding residual.

V. Conclusions

In this paper, we investigated a multi-channel multidimensional model for signals extracted from DNA gel electrophoresis. The proposed approach is based on using multi-capture imaging, which provides more information than using single-snapshot imaging or the finish-line method. Therefore, the obtained datasets provide high resolution which enables further signal processing for purposes of multidimensional time series analysis of the obtained signal. The multidimensional signal was aligned in order to increase resolution, then the resultant surface was fitted using an estimated list of average peak locations that were used to infer initial values for the multidimensional fit. The obtained simulation results has shown that the employed approach is successful for obtaining a synthetic form of the original dataset. Therefore, this method can be used as a start for future work, to produce enhanced versions using advanced signal processing techniques, e.g., accurate estimation of peak locations, distinguishing and omitting false peaks and robust models and reduction of background noise.

Acknowledgment

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) [EP/J015180/1].

J. R. Hopgood is funded by the Royal Academy of Engineering (RAE).

References

[1] V. G. Babskii, M. Y. Zhukov, and V. Yudovich, *Mathematical theory of electrophoresis*. Springer Science & Business Media, 2012.

[2] T. Schlick, *Molecular Modeling and Simulation: An Interdisciplinary Guide*. Springer Science & Business Media, 2013.

[3] H. Brenner, "Macrotransport processes," *Langmuir*, vol. 6, no. 12, pp. 1715–1724, 1990.

[4] E. R. Dougherty, A. Datta, and C. Sima, "Research issues in genomic signal processing," *IEEE Signal Processing Magazine,*, vol. 22, no. 6, pp. 46–68, 2005.

[5] V. Veljkovic, I. Cosic, D. Lalovic *et al.*, "Is it possible to analyze dna and protein sequences by the methods of digital signal processing?" *IEEE Transactions on Biomedical Engineering,*, no. 5, pp. 337–341, 1985.

[6] X.-Y. Zhang, F. Chen, Y.-T. Zhang, S. C. Agner, M. Akay, Z.-H. Lu, M. M. Y. Waye, and S. K.-W. Tsui, "Signal processing techniques in genomic engineering," *Proceedings of the IEEE*, vol. 90, no. 12, pp. 1822–1833, 2002.

[7] W. Wang and D. H. Johnson, "Computing linear transforms of symbolic signals," *IEEE Transactions on Signal Processing,*, vol. 50, no. 3, pp. 628–634, 2002.

[8] F. Parvaresh, H. Vikalo, S. Misra, and B. Hassibi, "Recovering sparse signals using sparse measurement matrices in compressed DNA microarrays," *SIEEE Journal of selected Topics in Signal Processing,*, vol. 2, no. 3, pp. 275–285, 2008.

[9] M. Buchner and S. Janjarasjitt, "Detection and visualization of tandem repeats in dna sequences," *IEEE Transactions on Signal Processing,*, vol. 51, no. 9, pp. 2280–2287, Sept 2003.

[10] N. Chakravarthy, A. Spanias, L. D. Iasemidis, and K. Tsakalis, "Autoregressive modeling and feature analysis of dna sequences," *EURASIP Journal on Applied Signal Processing*, vol. 2004, pp. 13–28, 2004.

[11] R. Lukac, K. N. Plataniotis, B. Smolka, and A. N. Venetsanopoulos, "A multichannel order-statistic technique for cdna microarray image processing," *IEEE Transactions on NanoBioscience,*, vol. 3, no. 4, pp. 272–285, 2004.

[12] E. Athanasiadis, D. Cavouras, P. P. Spyridonos, D. T. Glotsos, I. K. Kalatzis, G. C. Nikiforidis *et al.*, "Complementary dna microarray image processing based on the fuzzy gaussian mixture model," *IEEE Transactions on Information Technology in Biomedicine,*, vol. 13, no. 4, pp. 419–425, 2009.

[13] S. Samavi, S. Shirani, and N. Karimi, "Real-time processing and compression of dna microarray images," *IEEE Transactions on Image Processing,*, vol. 15, no. 3, pp. 754–766, 2006.

[14] J. R. Hopgood, "Improved resolution of chromatographic peak analysis using multi-snapshot imaging," in *IEEE 24th European Signal Processing Conference (EUSIPCO), in press,*, Budapest, Hungary, 2016.