

# DNA sequencing and parametric deconvolution

Lei Li  
Florida State University

July 10, 2001

## **Abstract**

One of the key practices of the Human genome project is Sanger DNA sequencing. Its data analysis part is called base-calling, which attempts to reconstruct target DNA sequences from fluorescence intensities generated by sequencing machines. In this paper, we present our modeling framework of DNA sequencing, in which a base-calling scheme arises naturally. A large portion of DNA sequencing errors come from the diffusion effect in electrophoresis, and deconvolution is the tool to solve this problem. We present a new version of the parametric deconvolution which is motivated by the spike-convolution model, and some recently obtained results regarding its asymptotics. One application of the asymptotics is to look at the resolution issue from the perspective of confidence intervals. We also report on an empirical study of the progressiveness of electrophoretic diffusion by way of estimating the slowly-changing width parameter in the spike-convolution model. Furthermore, we include an example of complete preprocessing of DNA sequencing data.

**Running title: DNA sequencing and parametric deconvolution**

**Key words and phrases: base-calling, color-correction, DNA sequencing, electrophoresis, parametric deconvolution, resolution, spike-convolution model, width.**

# 1 Introduction

The core genetic material of a human being consists of 23 pairs of chromosomes. The main part of each chromosome is a DNA molecule composed of two polymers called strands, each of which is a long sequence of four nucleotide bases — Adenine (A), Cytosine (C), Guanine (G) and Thymine (T) — supported by a sugar phosphate backbone. The two strands are complementary in the sense that A, G, C, and T on one strand pair with T, C, G, and A respectively on the other strand. One of the main goals of the Human Genome Project is to sequence the 3 billion or so base-pairs that make up the human chromosomes accurately. This blueprint is the basis of the genetic and genomic research that will eventually lead us to discover new diagnostic tools and treatments of human genetic diseases.

The current sequencing technique cannot sequence more than about 1000 base-pair at a time. Thus some kind of divide-and-conquer strategy is used to sequence the human genome. In the early stage of DNA sequencing history, the directed strategy was adopted by some genome centers. According to this strategy, we need to generate DNA clones from each chromosome at different resolution levels, and in turn produce many short fragments which can be sequenced directly. The fragments and clones should be generated in such a way that they cover the entire chromosome, or nearly so. After these fragments have been sequenced, we assemble and edit them, and then reconstruct the sequence of the chromosome. The work of generating DNA clones at different resolution levels and generating short fragments from clones is the task of physical mapping. With the sequencing cost recently dropping down dramatically, the shotgun strategy took over in the main sequencing projects. It skips some of the intermediate steps and directly generates many random fragments for sequencing. To ensure that these fragments cover the entire chromosome, a

larger fold of sequencing working load is required. An important mathematical model concerning the design and evaluation of these strategies was given by Lander and Waterman (1988). More results on physical mapping can be found in Waterman (1995). Yu and Speed (1997) studied the problem from an information theory perspective.

The need of modeling DNA sequencing emerges because its output, the genomic content, is becoming a cornerstone of life sciences. Although the completion of the first draft of the human genome was a historical milestone, a solid understanding of DNA sequencing will lead us to produce higher-quality genomes in the future and enhance the research along vertical directions such as genomic comparisons among individuals and across species. In this paper, we describe a modeling framework of DNA sequencing resulting from our research in the last few years. Based on existing knowledge, we propose a series of mathematical and probabilistic models to mimic the real sequencing procedure that incorporates cutting edge technologies of biology, chemistry, and physics. This model more or less enables us to simulate observations — DNA sequencing traces — from a target DNA sequence. Consequently, a scheme of base-calling which attempts to reconstruct the target sequence from observations arises naturally in consistency with the sequencing model. The significance of this framework is two-fold. On the one hand, it provides experimental researchers with some insights and guidance on how to improve the hardware components of sequencing such as enzyme designs and instruments. On the other hand, our model brings out the important issues of data analysis for the purpose of accurate base-calling. Like other researchers in the literature, we adopt a two-step strategy for base-calling: pre-processing followed by decision-making. To a great extent, we found that the accuracy of base-calling hinges on two key issues in preprocessing: deconvolution and color-correction, especially the former one. We developed a parametric deconvolution procedure, which is motivated by the so-called spike-convolution model, as a solution to the first

issue. In this paper, we present a new version of this procedure that includes a width parameter in the model, and some recently obtained results regarding its algorithms and asymptotics. One application of the asymptotics is to look at the resolution issue from the perspective of confidence intervals. We also report on an empirical study of the progressiveness of electrophoretic diffusion by way of estimating the slowly-changing width parameter in the spike-convolution model. We will discuss our work on color-correction briefly, and interested readers can find the references by following the given pointers.

We arrange the materials in this paper as follows. In Section 2 we describe DNA sequencing and our modeling framework. In Section 3 we describe the spike-convolution model and parametric deconvolution, and discuss several practical issues relating to DNA sequencing. Section 4 contains the mathematical proofs of some facts.

## **2 DNA sequencing and base-calling**

### **2.1 DNA sequencing**

Most currently used sequencing schemes are variants of the enzymatic method named after its inventor, Frederick Sanger. Each of these sequencing schemes consists of enzymatic reactions, electrophoresis, and some detection technique; see the book edited by Adams, Fields, and Venter (1994). First, four separate reactions are set up with a single-stranded DNA fragment annealed with an oligonucleotide primer. Each reaction contains the four normal precursors of DNA — that is, the deoxynucleotides dATP, dTTP, dCTP, and dGTP, together with the DNA polymerase as being used in the natural DNA replication. In addition, appropriate amounts of four dideoxy nucleotide terminators, ddATP, ddTTP, ddCTP, and ddGTP, are also present in the respective reactions. Thus people sometimes term it by dideoxy sequencing. In the ddA reaction containing

ddATP, polymers are extended from 5' to 3' end (the DNA strand is directed, and the two ends are named by 5' and 3' due to their chemical structures) by polymerase according to the template DNA, and the elongation of new strands is stopped once a ddATP is incorporated. Because the incorporation of ddATP is random, the ddA reaction should produce many copies of each possible sub-fragment starting with the same primer and ending with ddATP. Similarly, the ddG, ddC, and ddT reactions will produce many copies of each possible sub-fragment ending with ddGTP, ddCTP, and ddTTP respectively; see Russell (1995).

Next, electrophoresis is used to separate the DNA sub-fragments produced from the four reactions. DNA fragments are negatively charged in solution. If we load the DNA fragments into a slab gel or a gel-filled capillary and add an electric field, the fragments will move in the gel or capillary. The smaller the size of a fragment, the faster it runs through the electric field. In order to differentiate the four kinds of sub-fragments ending with ddGTP, ddCTP, and ddTTP, we can place them into four different but adjacent lanes in the gel. A more efficient color-coding strategy has been developed to permit sizing of all four kinds of DNA sub-fragments by electrophoresis in a single lane of a slab gel or in a capillary. That is, in each of the four reactions, the primers (or terminators) are labeled by one of four different fluorescent dyes. Laser-excited, confocal fluorescent detection systems are then used to excite the dyes in a region within the slab gel or capillary, and to collect and measure the fluorescence intensities emitted in four wavelength bands. These four fluorescence intensities are the raw data we can observe in practice. A segment of such kind of data — a four-component vector time series — is shown at the top of Figure 1. The four fluorescence intensities are not identical to the four dye concentrations passing through the detection region; rather, they are a transformed version of them. The four dye concentrations — another four-component vector time series — corresponding to the fluorescence intensities mentioned earlier

are shown in the middle of Figure 1, and they can be obtained by appropriate color-correction, which will be discussed more in Subsection 2.3.

The above Sanger sequencing procedure is schematically diagrammed in Figure 2. A hypothetical DNA fragment to be sequenced and its reverse complement are shown at the top. Please notice that we color-code the base A, G, C, and T respectively by red, black, green, and blue in all the figures of this paper. The four enzymatic reactions are illustrated in the middle. For the sake of simplicity, only one copy of each sub-fragment is presented. The hypothetical dye concentrations passing through the detection region are shown at the bottom. The laser device and detection system are skipped, and thus the fluorescence intensities are not shown in the figure.

## 2.2 DNA Base-calling

Base-calling is the analysis part of DNA sequencing, which attempts to reconstruct the target DNA sequence from the vector time series of fluorescence intensities. In Figure 2, some peaks of four colors are displayed. The rationale of base-calling is that each peak represents one base, and the order of peaks from the four channels is consistent with the order of nucleotide bases on the underlying DNA fragment. The hypothetical example in Figure 2 illustrates this process. Base-calling becomes harder for the data shown at the top of Figure 1, or for the data in the middle, if we focus on dye concentrations. The research in this area aims to make accurate and automated base-calling, along with appropriate assessment.

The dominating DNA sequencing devices being used are ABI sequencers produced by Applied Biosystems, Inc. Other producers include Beckman Coulter, Inc. etc. Some institutions use their own home-made devices for research of relatively small scales. ABI sequencers are accompanied by a base-calling software (1996). Several academic groups have also been conducting research on

base-calling. Those methods developed by Berno (1996), Berno and Stein (1995) at MIT (originally at Stanford), by Ives, Gesteland, and Stockham (1994) at University of Utah, and by Giddings, Brumley, Haker, and Smith (1993) at University of Wisconsin, adopted a similar framework, which consisted of two steps: preprocessing the data and decision-making. Preprocessing, which aims to clean up data, includes color correction, baseline subtraction, spacing adjustment, mobility shift adjustment, and peak sharpening. Decision-making is typically done by applying ad hoc algorithms to preprocessed data. Tibbetts (1994) treated the translation of sequencing images to DNA sequences as a pattern-recognition problem and used neural networks to call bases. The base-calling software *Phred*, developed by Ewing and Green (1998) and Ewing et al. (1998) at University of Washington, has an error rate smaller than that of ABI software as reported; see also Cawley (2000) for more comparison results. Our vision of carrying on the research in this regard is to make the model as clear as possible, for it provides a platform for further criticism and improvement. Nelson (1996) and Nelson and Speed (1996) provided an overview of this subject and described some initial efforts towards increasing base-calling accuracy and throughput by providing a rational, statistical model. This is the starting point of our modeling research in DNA sequencing.

### **2.3 A model framework of DNA sequencing and a strategy of base-calling**

Our strategy of base-calling is to first model the DNA sequencing to the best of our knowledge. That is, we examined each step of the physical DNA sequencing procedure, in which information of a DNA sequence is transformed from one form to another, and eventually into a vector time series—fluorescence intensities. A reasonable model should be able to simulate data similar to the real sequencing trace to some extent. Using such a model, we then can develop and optimize appropriate methods because the “artificial truth” can serve as a kind of reference.

We give a brief yet not necessarily complete account of the sources of uncertainties and complications intrinsic to DNA sequencing. As we can see later, they are the issues we need to face in base-calling. First, in the enzymatic reactions, the chance mechanism of replication and termination leads to our regarding the concentrations of the different DNA sub-fragments as random variables. Roughly speaking, this chance mechanism results in the variation of peak heights in the observations. It is also observed that the average peak heights decrease as time goes on. Second, the peak shape in the times series shown in the middle of Figure 1 or at the bottom of Figure 2 is a crucial factor for DNA sequencing, and this shape is referred to by point spread function (PSF) in spectroscopy. The point spread function is determined by the dynamics of polymers in electrophoresis, a complicated physical and chemical process. Some studies addressed this issue. A model using Brownian motion with drift results in an inverse Gaussian kernel function, with a scale parameter proportional to the square root of time; see Nelson (1996). A more delicate model, the reptation theory, results in an exponentially mediated Gaussian point spread function, which becomes wider and wider towards the end of electrophoresis; see Giddings (1965), Lumpkin, DeJardin and Zimm (1985), Luckey, Norris and Smith (1993). Other observed variability in spacing between peaks, peak width, mobility shifts of different dyes, temperatures, electronic field strength, and gel properties is rather experiment-specific. Scattered reports on interactions between bases are also found in the literature, but the issue is not among the primary considerations. Next in the data collection stage, dye concentrations are not observed directly by the detection system, as we mentioned earlier. Instead we measure fluorescence intensities emitted by the four dyes at four wavelength bands. Cross-talk comes in at this step because the emission spectra of the four dyes overlap. Finally, a slowly-changing baseline due to background fluorescence and other factors, and measurement errors, are also added into observations at this step.

We formulate the DNA sequencing procedure by a series of models, which are diagrammed at the left hand column of Figure 3. First, the sequence of the target DNA is encoded into a hidden Markov model (HMM), producing what we call the virtual signal containing four components. Different aspects of the HMM can be designed to incorporate variation in the concentrations of sub-fragments, the spacings between peaks, the spread of peaks, and the mobility shifts of the four dyes. We consider hidden Markov models because they have quite large modeling capacity and have dynamic programming type of algorithms to implement computations. By no means are they the only choice, and any machinery having enough modeling capacity and good algorithms is worth being considered. Second, the four components of the virtual signal are displaced with respect to one another according to the average mobility difference, resulting in the shifted virtual signal. Third, each component of the shifted virtual signal is convolved with a slowly-changing point spread function, to represent the average diffusion effect in electrophoresis. This convolved signal attempts to simulate the dye-labeled base concentrations traveling through the detection area in electrophoresis. Finally, these concentrations are further linearly transformed into fluorescence intensities, to approximate the cross-talk phenomenon. A slowly-changing baseline and white noise are added to the observations at this step to simulate measurement errors.

The above modeling of the information flow in DNA sequencing provides us with a natural framework for base-calling. That is, we undo each step in the model which mimics the real DNA sequencing, as is shown on the right hand column of Figure 3. Following the custom in the literature, we refer to these undoings by prefixing “de” to their corresponding mechanisms. Explicitly, we carry out de-cross-talk — color correction, to remove the dye effects, de-convolution to reconstruct our shifted virtual signal, de-mobility-shift to adjust for average mobility differences, and de-coding to make base calls. De-baseline — baseline subtraction — could either be done separately or

in combination with color correction. Other work such as de-noising and normalization may be needed depending on the methods being used, but are less important than the above issues. In the decoding stage, we could try either the Viterbi or marginal algorithm. However, the effectiveness of the decoding depends on the appropriateness of the HMM. Thus the determination of the HMM is a subtle problem. That is, we need to construct an appropriate hidden state space, design a topology of the transition pattern, and find estimates of the transition and output probabilities. Cawley (2000) in his thesis continued the research of hidden Markov model decoding using preprocessed data. As for the author of this article, most of the efforts have been devoted to color correction and deconvolution.

Several color correction algorithms were proposed in the literature such as Yin et al. (1996), Huang et al. (1997). To statisticians, the justification of an algorithm to a real problem remains unsolved until a model, in which assumptions could be verified to some degree, is established. Notice that both dye concentrations and the transformation representing the cross-talk phenomenon are to be estimated in the problem of de-cross-talk. In fact, without additional information this problem is ill-posed; or in statistical terms, the model is not identifiable. Li and Speed (1999) proposed a cross-talk model, and verified a crucial assumption in the model using data obtained from a specially designed experiment. That is, we placed the sub-fragments generated from the ddA, ddG, ddC, and ddT reactions into four different yet adjacent lanes of a slab gel. In this case the dye effects are restricted within each lane, and so the four dye concentrations can be observed. However, what we need from this experiment is nothing but the distribution of dye concentrations, for it is invariant with respect to different DNA sequence contents. With this piece of new information, the problem of de-cross-talk becomes well-posed. Consequently, an algorithm arose naturally from the model as a vehicle to achieve the goal of color correction. In Figure 1, the de-cross-talk was

carried out by our algorithm. Moreover, Kheterpal, Li, Speed, and Mathies (1998) found that the information contained in three fluorescence intensities is sufficient for reconstructing the four dye concentrations by using nonnegative least squares and a model selection procedure. This discovery brings more insights into the dye-based sequencing technique. For example, we proposed a new design to solve an even more challenging problem: sequence two DNA fragments in one lane — a first step towards high-order multiplex sequencing.

Once the data is properly color-corrected, we look at the problem of deconvolution. In Figure 1, sometimes we observe two or three peaks of the same color in a row. Four or more bases of the same kind are also observed in the genome, though their occurrence is relatively rare. Lack of caution in these cases would result in insertion and deletion errors of base-calling. Chen et al. (1992), Koop et al. (1993), and Lawrence and Solovyev (1994) reported that a large portion of DNA sequencing errors do come from these regions. In the next section, we present a new version of the so-called parametric deconvolution procedure, aiming to solve the above problem for DNA sequencing.

### 3 The spike-convolution model and parametric deconvolution

Let us first introduce some notation. We define the inner product of two functions  $y_1(t)$  and  $y_2(t)$  belonging to  $L^2[-\pi, \pi]$  by  $\langle y_1, y_2 \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} y_1(t) y_2(t) dt$ . For functions  $z_1(t)$  and  $z_2(t)$  well defined at the lattice points  $t_k = 2\pi k/n$ , where  $k = -[n/2], \dots, [(n-1)/2]$ , we also define the following inner product  $\langle z_1, z_2 \rangle_n = \frac{1}{n} \sum_{k=-[n/2]}^{[n/2]} z_1(t_k) z_2(t_k)$ . The norms induced by  $\langle \cdot, \cdot \rangle$  and  $\langle \cdot, \cdot \rangle_n$  are denoted by  $\| \cdot \|$  and  $\| \cdot \|_n$  respectively. We use the notation  $\xrightarrow{p}$  to represent convergence in probability. We also use  $G^T$  to represent the transpose of a matrix  $G$ . The material in this section is an expanded edition of the paper by Li and Speed (2000), hereafter referred to as

LS.

Our perception about the electrophoretic diffusion effect is represented by the spike-convolution model as defined below. We assume that the four kinds of sub-fragments, color-coded by four dyes, in a gel or capillary diffuse independently of each other. Therefore, we operate deconvolution for the four kinds of dye concentrations separately. It is conceived that the concentration of one fluorescence dye denoted by  $y(t)$  is the convolution of the virtual signal in Figure 3 and a point spread function  $w_\lambda$ . Namely,

$$y = w_\lambda * x, \quad (1)$$

where the point spread function,  $w_\lambda(t) = w(t/\lambda)$ , is generated from a prototype  $w(\cdot)$  and a scale parameter  $\lambda$ . We assume that the prototype of the point spread function is unchanged while its associated width parameter changes slowly over time, representing the progressive diffusion effect in electrophoresis. One strategy of deconvolution is to cut the sequencing traces into pieces in such a way that we can assume the width parameter  $\lambda$  is constant within each piece, and we adjust this parameter when moving from one piece to the next. Throughout this section, we assume that the width parameter can only take on values in a relatively small range:  $\lambda_0 < \lambda < \lambda_1$ , where  $\lambda_0$  and  $\lambda_1$  are two positive numbers. Let  $v_{\lambda,k} = \langle w_\lambda, e^{ikt} \rangle$  be  $w_\lambda$ 's Fourier coefficients. We assume

1.  $w_{\lambda_1}$  has finite support  $(-\kappa_1, \kappa_2)$ , where  $0 < \kappa_1, \kappa_2 < \pi$ ;
2.  $w(\cdot) \in C^2[-\pi, \pi]$ ;
3. for some integer  $K_0$  larger than the number of peaks in the unknown signal  $x$ ,  $v_{\lambda,k} \neq 0$ , where  $k = 0, \pm 1, \dots, \pm K_0$ .

Although the point spread function possibly has an infinite number of non-zero Fourier coefficients, the third condition required by the parametric deconvolution means that there is no vanishing

trigonometric moment before the index  $K_0$ . For the unknown signal  $x$  we proposed a specific form as follows.

$$x(t) = A_0 + \sum_{j=1}^p A_j \delta(t - \tau_j), \quad (2)$$

where  $\delta(\cdot)$  is the Dirac delta function, and the coefficients  $A_j$ , referred to by “heights” of the spikes, are positive. Thus the underlying signal  $x(t)$  is a linear combination of a finite number of spikes with positive heights, together with a constant baseline. We denote the signal  $x$  in (2) by  $SC(\delta; p; \mathbf{A}; \boldsymbol{\tau})$ , and refer to its convolution with  $w_\lambda$  as in (1) by  $SC(w_\lambda; p; \mathbf{A}; \boldsymbol{\tau})$ . We sample a  $SC(w_\lambda; p; \mathbf{A}; \boldsymbol{\tau})$  at the lattice points:  $\{2\pi k/n, k = -[n/2], \dots, [(n-1)/2]\}$ , add white noise to the signal, and generate

$$z(t_k) = y(t_k) + \epsilon(t_k) = A_0 + \sum_{j=1}^p A_j w(t_k - \tau_j) + \epsilon(t_k), \quad (3)$$

where the  $\{\epsilon(t_k)\}$  are i.i.d. with  $E(\epsilon(t_k)) = 0$ ,  $Var(\epsilon(t_k)) = \sigma^2$ , and a finite third moment. We use this model to formulate the diffusion effect and the measurement error mechanism of electrophoresis.

This setting leads us to the following idea of deconvolution: estimating the parameters in the spike-convolution model. The unknowns include the baseline; the error variance; the number, locations and heights of the spikes; and possibly the width parameter associated with the point spread function. The version without the width parameter can be found in Li (1998) and Li and Speed (2000). Notice that we define the signal in this model on a continuous scale, and we introduce a sparse positive Dirac delta train to represent occurrences of nucleotide bases. This leaves the room for a high resolution deconvolution.

### 3.1 Parametric deconvolution with a known width parameter

Following the general practice of statistical modeling, we first consider the identifiability issue.

**Proposition 3.1** *The spike-convolution model is identifiable when the width parameter is fixed. Namely, let  $y$  and  $\bar{y}$  be  $SC(w_\lambda; p; \mathbf{A}; \boldsymbol{\tau})$  and  $SC(w_\lambda; l; \bar{\mathbf{A}}; \bar{\boldsymbol{\tau}})$  respectively. Then  $\|y - \bar{y}\| > 0$  if the two sets of parameters are not identical.*

Next we consider the estimation problem. If we assume the measurement error is Gaussian and the number of spikes is known, then the maximum likelihood estimate or one-step estimate, as in the standard i.i.d. case, is asymptotically efficient; see LS. However, the maximization of the likelihood requires a good starting point, and this is a tough job because the likelihood surface is not unimodal even in the asymptotic sense. In addition, it is also desired to have a procedure which does not depend on the distributional assumption of measurement errors. Bearing these considerations in mind, we proposed a parametric deconvolution procedure. Because of the different roles played by the parameters in the model, it is of little hope to estimate them all in one step. The parametric deconvolution bundles up trigonometric moment estimates of the spike locations, least squares estimates of spike heights and baseline, and model selection techniques. The core of parametric deconvolution consists of two parts: model fitting and model selection. Keep in mind that we assume the width parameter is known in this subsection.

**Algorithm 3.1 Model-fitting.**

*Starting with the empirical trigonometric moments  $\hat{f}_k = \langle z, e^{ikt} \rangle_n$ , for any given nonnegative integer  $m \leq K_0$ , where  $K_0$  serves as an upper bound of the number of spikes, run the following steps.*

1. *Deconvolution: let  $\hat{g}_0 = \hat{f}_0$ ,  $\hat{g}_k = \hat{f}_k v_{\lambda,0}/v_{\lambda,k}$ , for  $k = \pm 1, \dots, \pm m$ .*
2. *Solving an eigen-value-vector problem: construct the Toeplitz matrix  $\hat{G}_m = (\hat{g}_{j-k})$ , and compute its smallest eigenvalue  $\hat{A}_0^{(m)}$  (assuming its multiplicity is one), and corresponding eigen-*

vector  $\hat{\alpha}^{(m)} = (\hat{\alpha}_0^{(m)}, \dots, \hat{\alpha}_m^{(m)})^T$ .

3. *Trigonometric moment estimates of spike locations: on the unit circle of the complex plane, find the  $m$  distinct roots of  $\hat{U}^{(m)}(z) = \sum_{j=0}^m \hat{\alpha}_j^{(m)} z^j$ , which we denote by  $\{e^{i\hat{\tau}_j^{(m)}}\}$ ,  $j = 1, \dots, m$ .*

4. *Eliminate those  $\{\hat{\tau}_j^{(m)}\}$  falling outside  $[-\pi + \kappa_1, \pi - \kappa_2]$ , and denote the locations of the remaining spikes by  $\{\bar{\tau}_j^{(\bar{m})}\}$ ,  $j = 1, \dots, \bar{m}$ , where  $\bar{m} \leq m$ .*

5. *Estimate the heights  $\bar{A}_j^{(\bar{m})}$  corresponding to these spikes by minimizing*

$$\| z(t) - \bar{A}_0^{(\bar{m})} - \sum_{j=1}^{\bar{m}} \bar{A}_j^{(\bar{m})} w(t - \bar{\tau}_j^{(\bar{m})}) \|_n^2 . \quad (4)$$

*This results in the least squares estimates of the baseline and heights.*

This algorithm outputs a  $SC(w_\lambda; \bar{m}; \bar{A}^{(\bar{m})}; \bar{\tau}^{(\bar{m})})$ . We make some notes on the implementation of this algorithm. First, algorithms of the Fourier transform and regression needed in step 1 and 5, respectively, have been well developed, and are not a problem at all. Second, we only need to calculate the smallest eigenvalue  $A_0$  and its eigenvector of the Toeplitz matrix  $G$ , which are the same as the largest eigenvalue and its eigenvector of the inverse matrix  $G^{-1}$ . As a matter of fact, there is a nice solution to this problem. On the one hand, for the Toeplitz matrix the inverse can be calculated using the Trench algorithm, which requires only  $O(N^2)$  flops; see Golub and Van Loan (1996). On the other hand, for a symmetric matrix the largest eigenvalue and its eigenvector can be computed very quickly by the the power method. That is, we generate a sequence  $\{A_{0,\{k\}}, \alpha_{\{k\}}\}$  using the following iteration.

$$\begin{cases} \beta_{\{k\}} &= G \alpha_{\{k-1\}}, \\ \alpha_{\{k\}} &= \beta_{\{k-1\}} / \|\beta_{\{k-1\}}\|_2, \\ A_{0,\{k\}} &= \alpha_{\{k\}}^T G \alpha_{\{k\}}, \end{cases}$$

where  $\|\cdot\|_2$  is the Euclidean norm of a vector. It can be proved that the sequence converges to  $\{A_0, \alpha\}$  at an exponential rate if the smallest eigenvalue is simple — namely, its multiplicity is one; see Riesz and Nagy (1955) or Golub and Van Loan (1996). If the multiplicity is larger than one, then we may observe the so-called “wobbly” phenomenon. Theoretically, this phenomenon is not really a problem for parametric deconvolution according to the argument following Proposition 5.1 in LS. Numerically, we have not encountered this problem in analyzing real sequencing data. Finally, the polynomial to be solved in step 3 involves complex variables. In general, solving a polynomial of a complex variable is not an easy one, for we have to search through the unbounded complex plane. Surprisingly, it is found that we can regard it as an equation of a real variable defined on  $[-\pi, \pi]$  because the following result implies that all the roots of  $\sum_{j=0}^m \hat{\alpha}_j^{(m)} z^j$  reside strictly on the unit circle.

**Theorem 3.1** 1. Given a  $SC(\delta; m; \mathbf{A}; \boldsymbol{\tau})$ , let  $G_m$  be the Toeplitz matrix constructed from its Fourier coefficients  $\{g_k\}$ . Write  $U^{(m)}(z) = \prod_{j=1}^m (z - e^{i\tau_j}) = \sum_{j=0}^m \alpha_j z^j$ . Then  $A_0$  is the smallest eigenvalue of  $G_m$ . Its multiplicity is one and its eigenvector is  $(\alpha_0, \dots, \alpha_m)^T$ . The  $\{A_j\}$  satisfy the following linear system:

$$\frac{1}{2\pi} \begin{pmatrix} 1 & 1 & \dots & 1 \\ e^{i\tau_1} & e^{i\tau_2} & \dots & e^{i\tau_m} \\ \vdots & \vdots & \ddots & \vdots \\ e^{i(m-1)\tau_1} & e^{i(m-1)\tau_2} & \dots & e^{i(m-1)\tau_m} \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_m \end{pmatrix} = \begin{pmatrix} g_0 - A_0 \\ g_1 \\ \vdots \\ g_{m-1} \end{pmatrix}. \quad (5)$$

2. Conversely, suppose we are given  $m+1$  complex numbers  $\{g_j, 0 \leq j \leq m\}$ , let  $g_{-j} = \overline{g_j}$  for  $1 \leq j \leq m$ . Assume that the smallest eigenvalue  $A_0$  of the Toeplitz matrix  $G_m = (g_{j-i})_{i,j=0,\dots,m}$  is simple. Let the smallest eigenvector be  $\alpha = (\alpha_0, \dots, \alpha_m)^T$ , and  $U^{(m)}(z) = \sum_{j=0}^m \alpha_j z^j$ . Then there exists a unique  $SC(\delta; m; \mathbf{A}; \boldsymbol{\tau})$  whose first  $m+1$  Fourier coefficients are  $\{g_j, 0 \leq j \leq m\}$ . The  $\{\tau_j\}$  are determined from the  $m$  distinct roots  $\{e^{i\tau_j}\}$  of  $U^{(m)}(z)$  lying exactly on the unit

circle. The  $\{A_j\}$  are determined by the linear system (5), and the resulting heights are positive.

This result is the heart and soul of the parametric deconvolution method. Thus, we detail one proof here in order to illuminate the structure of the spike-convolution model. The first part is easy to check. We note that  $\alpha = (\alpha_0, \dots, \alpha_m)^T$  is conjugate symmetric except a constant of modulus one. We can show this through the reverse operator  $J$  defined as

$$J_m = \begin{pmatrix} 0 & \dots & 0 & 1 \\ 0 & \dots & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \dots & 0 & 0 \end{pmatrix}.$$

Notice that  $J_m G_m J_m J_m \alpha = A_0 J_m \alpha$ , and  $\overline{J_m G_m J_m J_m \alpha} = A_0 \overline{J_m \alpha}$ , and thus  $G_m \overline{J_m \alpha} = A_0 \overline{J_m \alpha}$ . By the uniqueness of the eigenvector, we know  $\alpha = \overline{J_m \alpha}$  except a constant of modulus 1. Because of this property, if  $z_0$  is a root of  $K(z)$ , then  $\overline{z_0^{-1}}$  is also a root of  $K(z)$ .

As for the second part, we here gives a measure-theoretic proof. We can regard the positive spikes in the model as a kind of energy — point masses — generated by the trains of nucleotide bases. The distributions of these point masses of the four components characterize the target DNA sequence.

First we prove that we can assume  $\alpha_0 = \alpha_m = 1$ . Otherwise, without loss of generality, say  $\alpha_0 = \alpha_m = 0$  but  $\alpha_1 = \overline{\alpha_{m-1}} \neq 0$ , then by the structure of Toeplitz matrix, we have

$$\overline{(\alpha_1, \dots, \alpha_{m-1}, 0, 0)}^T G_m = A_0 \overline{(\alpha_1, \dots, \alpha_{m-1}, 0, 0)}^T.$$

Thus  $(\alpha_1, \dots, \alpha_{m-1}, 0, 0)^T$  is another eigenvector corresponding to  $A_0$ . This contradicts the assumption that  $A_0$  has a multiplicity of one. Let  $\tilde{g}_j = g_j$ ,  $j = \pm 1, \dots, \pm m$ ,  $\tilde{g}_0 = g_0 - A_0$ , and construct Toeplitz matrices  $\tilde{G}_m = (\tilde{g}_{j-i})_{i,j=0,\dots,m}$  as usual. It is obvious  $\tilde{G}_m \geq 0$ . Its smallest

eigenvalue is 0 and simple, and the corresponding eigenvector is  $\alpha = (\alpha_0, \dots, \alpha_m)^T$ . For  $k > m$ , let

$$\tilde{g}_k = - \sum_{j=0}^{m-1} \alpha_j \tilde{g}_{k-m+j} = -(\alpha_0 \dots \alpha_{m-1}) \begin{pmatrix} \tilde{g}_{k-m} \\ \vdots \\ \tilde{g}_{k-1} \end{pmatrix} \quad (6)$$

and for  $k < -m$ , let  $\tilde{g}_k = \overline{\tilde{g}_{-k}}$ . This implies that for any  $k \geq 0$ , we have

$$\begin{pmatrix} \overline{\alpha_0} & \overline{\alpha_1} & \cdots & \overline{\alpha_m} & 0 & \cdots \\ 0 & & & & & \\ \vdots & & & & & \\ 0 & & & I_{m+k+1} & & \\ 0 & & & & & \\ \vdots & & & & & \end{pmatrix} \tilde{G}_{m+k+1} \begin{pmatrix} \alpha_0 & 0 & \cdots & 0 & 0 & \cdots \\ \alpha_1 & & & & & \\ \vdots & & & & & \\ \alpha_m & & & I_{m+k+1} & & \\ 0 & & & & & \\ \vdots & & & & & \end{pmatrix} = \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 & \cdots \\ 0 & & & & & \\ \vdots & & & & & \\ 0 & & & \tilde{G}_{k+m} & & \\ 0 & & & & & \\ \vdots & & & & & \end{pmatrix},$$

where  $I_{m+k+1}$  is the identity matrix of order  $m+k+1$ . By induction, we can see that  $\tilde{G}_{m+k} \geq 0$ , for any  $k > 0$ . Thus by the Herglotz Theorem, there exists a measure  $dF$  on  $[-\pi, \pi]$  such that  $\tilde{g}_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ikt} dF(t)$ . Now we decompose  $F(t)$  into two parts  $F = F^a + F^s$ , where  $F^a$  is the absolute continuous part with respect to Lebesgue measure while  $F^s$  is the singular part. Notice that

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} |K(e^{it})|^2 dF(t) = \overline{\alpha} \tilde{G}_m \alpha = 0. \quad (7)$$

Thus

$$\int_{-\pi}^{\pi} |K(e^{it})|^2 dF^a = 0, \quad (8)$$

$$\int_{-\pi}^{\pi} |K(e^{it})|^2 dF^s = 0. \quad (9)$$

We can write  $dF^a(t) = f(t)dt$ , so  $\int_{-\pi}^{\pi} |K(e^{it})|^2 f(t) dt = 0$ . Since  $|K(e^{it})|^2$  has at most finite zeros,  $f(t) = 0$  almost everywhere, which implies  $dF^a = 0$ . Next it is inferred from (9) that  $dF^s$  has nonzero masses at  $m'$  points, where  $m' \leq m$ , since  $K(z)$  has  $m$  zeros. Furthermore, we conclude  $m' < m$  is impossible. Otherwise, by the first half of the theorem, there exists a  $\beta = (\beta_0 \cdots \beta_{m'})^T$  such that  $\tilde{G}_{m'} \beta = 0$ . So  $(\beta_0 \cdots \beta_{m'} 0 \cdots 0)^T$  is another eigenvector of  $\tilde{G}_m$  corresponding to the eigenvalue 0. This contradicts the assumption that the multiplicity of the eigenvalue 0 is one.

Hence  $F$  is a discrete measure with masses  $\{A_j, A_j > 0\}$  respectively at  $\{-\pi < \tau_1 < \dots < \tau_m < \pi\}$ . Next by (7), it is obvious  $K(e^{i\tau_j}) = 0$ , for  $j = 1, \dots, m$ . Thus all the roots of  $K(z)$  reside on the unit circle. The rest can be easily checked. The generalization of this result from  $SC(\delta; m; \mathbf{A}; \boldsymbol{\tau})$  to  $SC(w_\lambda; m; \mathbf{A}; \boldsymbol{\tau})$  is Theorem 3.2 in LS, and is the direct justification of Algorithm 3.1..

In Algorithm 3.1, take  $m = p$ . With a little abuse of notation, we use  $\hat{\boldsymbol{\tau}} = \{\hat{\tau}_j, j = 1, \dots, p\}$  to denote the the trigonometric moment estimates of the spike locations obtained in step 1, 2, 3. With probability tending to one they will not fall into the boundary region and be eliminated in step 4 because of their consistency. This conclusion was proved in Theorem 3.3 in LS, where we also gave the central limit theorem for the trigonometric moment estimates of the spike locations, heights, and baseline. The trigonometric moment estimates of the spike heights and baseline are given by the following Vandermonde linear system:

$$v_{\lambda,0} \begin{pmatrix} 1 & 1 & \dots & 1 \\ e^{i\hat{\tau}_1} & e^{i\hat{\tau}_2} & \dots & e^{i\hat{\tau}_p} \\ \vdots & \vdots & \ddots & \vdots \\ e^{i(p-1)\hat{\tau}_1} & e^{i(p-1)\hat{\tau}_2} & \dots & e^{i(p-1)\hat{\tau}_p} \end{pmatrix} \begin{pmatrix} \tilde{A}_1 \\ \tilde{A}_2 \\ \vdots \\ \tilde{A}_p \end{pmatrix} = \begin{pmatrix} \hat{g}_0 - \hat{A}_0^{(p)} \\ \hat{g}_1 \\ \vdots \\ \hat{g}_{p-1} \end{pmatrix},$$

where  $(\tilde{A}_1, \dots, \tilde{A}_p)$  are the trigonometric moment estimates of  $(A_1, \dots, A_p)^T$ . It can be inferred from Theorem 3.1 that  $(\tilde{A}_1, \dots, \tilde{A}_p)$  are positive. An efficient algorithm requiring only  $O(N^2)$  flops exists to solve this Vandermonde linear system; see Golub and Van Loan (1996). However, the least squares method adopted in Algorithm 3.1 has the following attractive asymptotics. Write  $\Xi_{\boldsymbol{\tau}} = (\xi_{\tau_0}, \xi_{\tau_1}, \dots, \xi_{\tau_p})^T$ , where the components  $\xi_{\tau_0} = 1$ ,  $\xi_{\tau_j} = w(t - \tau_j)$ ,  $j = 1, \dots, p$  are functions defined on  $[-\pi, \pi]$ . Then the least squares estimates of the baseline and spike heights are given by

$$(\hat{A}_0, \hat{A}_1, \dots, \hat{A}_p) = \langle \Xi_{\hat{\boldsymbol{\tau}}}, \Xi_{\hat{\boldsymbol{\tau}}}^T \rangle_n^{-1} \langle \Xi_{\hat{\boldsymbol{\tau}}}, z \rangle_n, \quad (10)$$

where  $\langle \Xi_{\hat{\boldsymbol{\tau}}}, \Xi_{\hat{\boldsymbol{\tau}}}^T \rangle_n = [\langle \xi_{\hat{\tau}_j}, \xi_{\hat{\tau}_k} \rangle_n]_{j,k=0,\dots,p}$ ,  $\langle \Xi_{\hat{\boldsymbol{\tau}}}, z \rangle_n = (\langle \xi_{\hat{\tau}_1}, z \rangle_n, \dots, \langle \xi_{\hat{\tau}_p}, z \rangle_n)^T$ . Please remember that the inner products  $\langle \cdot, \cdot \rangle$ ,  $\langle \cdot, \cdot \rangle_n$  are those defined at the beginning of this

section.

**Proposition 3.2**  $(\hat{A}_0, \hat{A}_1, \dots, \hat{A}_p)$  are consistent estimates; moreover, they are asymptotically normally distributed with zero mean and variance  $\sigma^2 < \Xi_{\mathcal{T}}, \Xi_{\mathcal{T}}^T >^{-1}$ .

In this asymptotic sense, the least squares estimates of baseline and spike heights based on the trigonometric moment estimates of the spike locations perform as well as if the parameter values of the spike locations were known. Therefore, we expect that the least squares estimates outperform the trigonometric moment estimates of the baseline and spike heights. Indeed, this performance has been observed in our simulation study. More generally, Proposition 3.2 holds as long as we have a set of consistent estimates of the spike locations regardless of their efficiency.

Algorithm 5.2 in LS serves as the model selection procedure in the parametric deconvolution. Unlike the usual practice of model selection, this two-stage procedure has dual purposes: estimate the model order and help to generate a set of estimates of spike locations with smaller bias and variance. The simulation study in LS showed that the bias and variance of the direct trigonometric moment estimates are much larger than those obtained from this model selection procedure under the Gaussian assumption, when the model order is assumed to be known. Our strategy to achieve the goal is to find a "best over-fitting" model in the first stage and eliminate the false spikes in the second stage. First, let us establish the following fact.

**Proposition 3.3** *In the first stage of Algorithm 5.2 in LS, the order will not be under-estimated with probability tending to one. Namely,  $P(\bar{m}_0 \geq p) \rightarrow 1$ .*

How does this two-stage model selection procedure complement the model fitting procedure to offer a reasonably good solution to the parameter estimation problem? This is an interesting yet challenging theoretical problem. Proposition 3.3 shows that the locations in the model

$SC(w_\lambda; \bar{m}_0; \bar{\mathbf{A}}^{(\bar{m}_0)}; \bar{\boldsymbol{\tau}}^{(\bar{m}_0)})$ , which is selected from the first stage, is constructed from the Toeplitz matrix  $\hat{G}_{\bar{m}_0} = (\hat{g}_{j-k})$ , where  $\bar{m}_0 \geq p$  in the probability sense. Proposition 5.1 in LS, together with the heuristic following it, imply that the spike locations obtained in this way,  $\bar{\tau}_1^{(\bar{m}_0)}, \dots, \bar{\tau}_{(\bar{m}_0)}^{(\bar{m}_0)}$ , contain a subset that are close to the true spike locations if the sample size is large enough. The second stage of the model selection is essentially a backward deletion procedure. As shown by An and Gu (1985), the backward deletion procedure is generally consistent. Thus we expect that any false spikes in  $\bar{\tau}_1^{(\bar{m}_0)}, \dots, \bar{\tau}_{(\bar{m}_0)}^{(\bar{m}_0)}$  could be deleted in the second stage of Algorithm 5.2 in LS, and this would result in an consistent estimate of the model order. Next, Proposition 3.2 shows that once we have a set of consistent estimates of the spike locations, the baseline and spike heights can be estimated as well as if the spike locations were known. The picture looks quite nice when these propositions are combined. Yet we still cannot provides an analytical interpretation of the phenomenon observed in our simulation: the desired subset of  $\bar{\tau}_1^{(\bar{m}_0)}, \dots, \bar{\tau}_{(\bar{m}_0)}^{(\bar{m}_0)}$  are fairly good estimates of the true spike locations in terms of bias and variance; see Example 6.1 and Table 1 in LS.

### 3.2 Adjusting the unknown width parameter

In this subsection, we regard the width parameter  $\lambda$  as part of the parameters to be estimated. Remember that this width parameter in the case of DNA sequencing, probably in other cases too, takes values only in a narrow range. The identifiability of the spike-convolution model with fixed width parameter is established in Proposition 3.1. But life becomes more complicated when the free width parameter is included. We assume throughout this subsection that the identifiability remains valid in a local neighborhood of the true model. This assumption avoids technical complications while it is reasonable for the DNA sequencing data.

We first consider the likelihood method by assuming that the measurement errors are i.i.d. Gaussian. The  $-2\log\text{likelihood}$  of the observations generated from the model is given by

$$n \log(2\pi\sigma^2) + \frac{1}{\sigma^2} \sum_l \left\{ z(t_l) - A_0 - \sum_{j=1}^p A_j w_\lambda(t_l - \tau_j) \right\}^2. \quad (11)$$

We write  $\theta = (\lambda, A_0, A_1, \dots, A_p, \tau_1, \dots, \tau_p)^T$ , and sometimes we use  $y_\theta(t)$  to denote  $SC(w_\lambda; p; \mathbf{A}; \boldsymbol{\tau})$ .

Write the gradient vector with respect to  $\theta$  by  $\nabla_\theta = (\partial \log L / \partial \theta)^T$ . The Fisher information matrix is, as usual, defined by  $I_\theta = \frac{1}{n} E[\nabla_\theta \nabla_\theta^T]$ . Then the following can be checked.

**Proposition 3.4** *Let  $\Psi_\theta = (\psi_\lambda, \psi_{A_0}, \psi_{A_1}, \dots, \psi_{A_p}, \psi_{\tau_1}, \dots, \psi_{\tau_p})^T$ , where*

$$\psi_\lambda = \sum_{j=1}^p A_j (t - \tau_j) w'_\lambda(t - \tau_j), \quad \psi_{A_0} = 1, \quad \psi_{A_j} = w_\lambda(t - \tau_j), \quad \psi_{\tau_j} = -A_j w'_\lambda(t - \tau_j), \quad j = 1, \dots, p.$$

*Then  $I_\theta = \frac{1}{\sigma^2} \int_{-\pi}^{\pi} [\Psi_\theta(t) \Psi_\theta(t)^T] dt$ .*

We can use the Gauss-Newton method — see Algorithm 4.1 in LS — to adjust the parameter estimates. Though this procedure can be iterated, a similar asymptotic result as Theorem 4.2 in LS shows one step is enough for the consideration of efficiency. The above procedure of maximizing the likelihood is equivalent to minimizing the residual sum of squares. Even if we drop the Gaussian assumption, we still can use  $L^2$  norm as our loss function. We found the following simple method is quite effective in tuning the width parameter for DNA sequencing trace.

### Algorithm 3.2 Width Tuning

*Generate a set of lattice points in the range  $(\lambda_0, \lambda_1)$ . For each of these values of  $\lambda$ , apply Algorithm 3.1 and Algorithm 5.2 in LS to fit the data, and compute the residual sum of squares. Choose the value of  $\lambda$  that minimizes the residual sums of squares.*

### 3.3 Discussion and examples

Now we apply the parametric deconvolution procedure to the color-corrected data shown in the middle of Figure 3. Bear in mind that we operate deconvolution for each dye concentration separately. In the spike-convolution model, we assume there are no spikes near the two ends. In practice, we cut the trace of one dye concentration into pieces at the valley points in such a way that each piece has room for about 12-20 bases. We then scale each piece to the range  $[-\pi, \pi]$ , apply the parametric deconvolution procedure — Algorithm 3.1 and Algorithm 5.2 in LS — to it, and get the output back to the original scale. Because these pieces are not necessarily of exactly the same length, the width parameters in their corresponding spike convolution models vary slightly from one piece to another, even though they are constant in the original scale. From now on throughout this subsection, the width parameter will be meant in the original scale. We apply Algorithm 3.2 to the four dye concentrations shown in Figure 3, and obtained the estimates of their width parameters. We found that they are approximately the same. That is, the loss in terms of residual sum of squares is negligible by assuming the width parameter is constant across the four components. This is equivalent to say that the electrophoretic diffusion effects of the four dyes progress in the same pace. At the bottom of Figure 3, the output of the parametric deconvolution — spikes — is depicted in comparison to the raw data, and color-corrected data. There are 49 or so nucleotide bases in this window, and each dye component is chopped into two or three pieces. The width parameter is taken to be 4.03 across the time and across the four components. The spike locations are rounded off to the closest integers. It is noticed that all consecutive bases are well separated. Correct base-calling can be even made by setting a proper threshold.

In the literature, the limitation and capability of peak separation is described by the concept —

resolution. According to Grossman et al. (1992) or Luckey et al. (1993), this is defined as follows in the case of Gaussian.

$$Resolution = \frac{\tau_2 - \tau_1}{2\sqrt{2\log 2}\lambda}, \quad (12)$$

where  $\tau_1$  and  $\tau_2$  are the centers of two adjacent Gaussians, and  $\lambda$  is the standard deviation. If we assume the two Gaussians have the same heights, then the two peaks will merge into one when the resolution falls below 0.5. Thus they are indistinguishable by naive bump-hunting. The situation is even worse if the two Gaussians' heights are not identical. This simple view of resolution does not take into account of any measurement errors, general point spread functions, and general spike configurations potentially with more than two spikes. The spike-convolution model provides us with another perspective to study the issue. According to the asymptotics of the one-step estimate, see Theorem 4.1 in LS, we can construct the confidence intervals for the spike locations and heights.

**Proposition 3.5** *Let the diagonal of the inverse of  $\int_{-\pi}^{\pi} [\Psi_{\theta}(t)\Psi_{\theta}(t)^T] dt$  be  $\rho$ ; see Proposition 3.4. Let the one-step estimate of the parameter vector be  $\theta_{new}$ . Then we have  $100(1-\alpha)\%$  asymptotically simultaneous confidence intervals:  $\theta_{new} \pm z_{\frac{\alpha}{2}}\sigma_{new} \rho_{\theta_{new}}/\sqrt{n}$ , where  $z_{\frac{\alpha}{2}}$  is the  $1 - \frac{\alpha}{2}$  quantile of the standard normal distribution,  $\sigma_{new}$  is the one-step estimate of  $\sigma$ .*

In light of this result, we take a new look at the resolution issue: we can tell apart two adjacent spikes at the confidence level  $100(1 - \alpha)\%$  if the following two items are satisfied:

1. the confidence intervals of their locations do not overlap;
2. the confidence lower bounds of their heights are positive.

This characterizes the capability and limitation of the parametric deconvolution — a model-based procedure. It is interesting to notice the interplay of the magnitude of the measurement error, the spike configuration, and the point spread function including the width parameter in this treatment.

To a great extent the width parameter  $\lambda$  determines the resolution in DNA sequencing, for the change in the inter-arrival time is small compared with that of the width in a local range. The width value depends on the experimental factors such as gel type, electric field strength, and temperature. The width value also depends on the position of the nucleotide base relative to the primer because diffusion becomes stronger as electrophoresis progresses. In order to study how the width evolves, we look at a larger segment of sequencing data than that shown in Figure 1. The data set starts with the 10-th base from the primer and include about 5000 scans. There are more than 500 nucleotide bases in this range, for on average there are about 8 to 9 scans between adjacent bases. We color correct the data and choose one dye concentration for further study. We did so because the four dye concentrations share a similar width pattern, as we mentioned earlier. We cut the data into pieces, each consisting of about 100 to 150 scans. We estimate the width parameters for a few pieces, including the two at the ends, by manually checking the deconvolution results and residual sum of squares. Then we interpolate the width across the whole range. This is the starting point for further width tuning. For each piece, we add the two pieces just before it and after it. This process create a window consisting of about 300 to 450 scans. Then we tune the width parameter by minimizing the residual sum of squares in a neighborhood centering around the starting value, and letting it be the estimate of width for the center piece. According to this scheme, adjacent windows overlap by about two thirds. By doing so, we can obtain a more accurate estimate of the width parameter for that region. By setting a bounded neighborhood for the parameter, we can avoid un-identifiability problem. The analysis in Grossman et al. (1992) and Luckey et al. (1993) implies that the square of the width parameter is proportional to the time. In Figure 4, we plot the square of the fitted width parameter versus the time. A straight line is fitted to the scatter plot by the least squares method. A more or less linear trend can be seen,

except for the widths corresponding to nucleotide chains of small sizes.

Most existing deconvolvers such as Jansson's method are non-parametric type in nature, because they do not assume a specific form for the unknown signal. Jansson's method has been widely used in spectroscopy. It demands very little in computation and provides a reasonably good solution in many cases. Li and Speed (2001) compared parametric deconvolution, Jansson's method, and the deconvolver which minimizes a Kullback-Leibler divergence, by both analysis and numerical examples. The results on parametric deconvolution are quite encouraging, and it seems modeling can indeed improve the data preprocessing in DNA sequencing by a great deal. It is our hope that this new perception will benefit researchers in other scientific areas as well.

## 4 Appendix

**Proof of Proposition 3.1:** Let  $x(t)$  and  $\bar{x}(t)$  be the  $SC(\delta; p; \mathbf{A}; \boldsymbol{\tau})$  and  $SC(\delta; l; \bar{\mathbf{A}}; \bar{\boldsymbol{\tau}})$  corresponding to  $y$  and  $\bar{y}$  respectively. Their Fourier coefficients are denoted by  $\{g_k\}$ ,  $\{\bar{g}_k\}$ . According to the convolution theorem, the Fourier coefficients of  $y$  are  $f_k = g_k v_{\lambda,k}$ , and those of  $\bar{y}$  are  $\bar{f}_k = \bar{g}_k v_{\lambda,k}$ . If  $\|y - \bar{y}\| = 0$ , then the Fourier coefficients of  $y$  and  $\bar{y}$  are identical. Consequently, we have  $\{g_k\} = \{\bar{g}_k\}$  for  $0 \leq k \leq K_0$ . Without loss of generality, we assume  $p \geq l$ . According to Theorem 3.1, the parameter values in  $SC(\delta; p; \mathbf{A}; \boldsymbol{\tau})$  is uniquely determined from the smallest eigenvalue and its eigenvector of the Toeplitz matrix  $(g_{j-k})_{k,j=0,\dots,m}$ , and so is  $SC(\delta; l; \bar{\mathbf{A}}; \bar{\boldsymbol{\tau}})$ . This contradicts the assumption that  $SC(w_\lambda; p; \mathbf{A}; \boldsymbol{\tau})$  and  $SC(w_\lambda; l; \bar{\mathbf{A}}; \bar{\boldsymbol{\tau}})$  are different.

Later we need the following lemma.

**Lemma 4.1** *If  $\hat{A} \xrightarrow[p]{} A$ , and  $\hat{\tau} \xrightarrow[p]{} \tau$ , then  $\|\hat{A} w_\lambda(t - \hat{\tau}) - A w_\lambda(t - \tau)\|_n \xrightarrow[p]{} 0$ .*

The Taylor expansion  $\hat{A} w_\lambda(t - \hat{\tau})$  around  $A w_\lambda(t - \tau)$ , and the boundedness of  $w'_\lambda(t)$  gives

$$|\hat{A} w_\lambda(t - \hat{\tau}) - A w_\lambda(t - \tau)| \leq M_1 |\hat{A} - A| + M_2 |\hat{\tau} - \tau|,$$

where  $0 < M_1, M_2 < \infty$ , and do not depend on  $t$ . Then the conclusion follows true.

**Proof of Proposition 3.2:** First notice that

$$\begin{aligned} & \langle \xi_{\hat{\tau}_j}, \xi_{\hat{\tau}_k} \rangle_n - \langle \xi_{\tau_j}, \xi_{\tau_k} \rangle_n = \langle \xi_{\hat{\tau}_j}, \xi_{\hat{\tau}_k} - \xi_{\tau_k} \rangle_n + \langle \xi_{\hat{\tau}_j} - \xi_{\tau_j}, \xi_{\tau_k} \rangle_n \\ & \leq \|\xi_{\hat{\tau}_j}\|_n \|\xi_{\hat{\tau}_k} - \xi_{\tau_k}\|_n + \|\xi_{\hat{\tau}_j} - \xi_{\tau_j}\|_n \|\xi_{\tau_k}\|_n \xrightarrow{p} 0, \end{aligned}$$

where we apply the Cauchy-Schwarz inequality to the second last step, and Lemma 4.1 to the last step. Hence we have  $\langle \Xi_{\hat{\boldsymbol{\tau}}}, \Xi_{\hat{\boldsymbol{\tau}}}^T \rangle_n \xrightarrow{p} \langle \Xi_{\boldsymbol{\tau}}, \Xi_{\boldsymbol{\tau}}^T \rangle_n$ . Next notice that

$\langle \xi_{\hat{\tau}_j} - \xi_{\tau_j}, \epsilon \rangle_n \leq \|\xi_{\hat{\tau}_j} - \xi_{\tau_j}\|_n \|\epsilon\|_n \xrightarrow{p} 0$  because  $\|\xi_{\hat{\tau}_j} - \xi_{\tau_j}\|_n \xrightarrow{p} 0$  according to Lemma 4.1, and  $\|\epsilon\|_n \xrightarrow{p} \sigma^2$  according to the law of large numbers. Therefore we have  $\langle \Xi_{\hat{\boldsymbol{\tau}}} - \Xi_{\boldsymbol{\tau}}, \epsilon \rangle_n \xrightarrow{p} 0$ .

Similarly we have  $\langle \Xi_{\hat{\boldsymbol{\tau}}} - \Xi_{\boldsymbol{\tau}}, y \rangle_n \xrightarrow{p} 0$ ; hence

$$\langle \Xi_{\hat{\boldsymbol{\tau}}} - \Xi_{\boldsymbol{\tau}}, z \rangle_n = \langle \Xi_{\hat{\boldsymbol{\tau}}} - \Xi_{\boldsymbol{\tau}}, y \rangle_n + \langle \Xi_{\hat{\boldsymbol{\tau}}} - \Xi_{\boldsymbol{\tau}}, \epsilon \rangle_n \xrightarrow{p} 0.$$

These results together with the following decomposition

$$\langle \Xi_{\hat{\boldsymbol{\tau}}}, z \rangle_n = \langle \Xi_{\boldsymbol{\tau}}, z \rangle_n + \langle \Xi_{\hat{\boldsymbol{\tau}}} - \Xi_{\boldsymbol{\tau}}, z \rangle_n,$$

allow us to apply the Slutsky Theorem to the least squares estimates in (10), and come to the conclusion that  $(\hat{A}_0, \hat{A}_1, \dots, \hat{A}_p)$  has the same asymptotic distribution as  $\langle \Xi_{\boldsymbol{\tau}}, \Xi_{\boldsymbol{\tau}}^T \rangle_n^{-1} \langle \Xi_{\boldsymbol{\tau}}, z \rangle_n$ , in which the spike locations are assumed to be known. Then an application of the Lindeberg-Feller theorem for triangular arrays — see Durrett (1991) — tells us that it is a normal distribution with zero mean, and variance  $\sigma^2 \langle \Xi_{\boldsymbol{\tau}}, \Xi_{\boldsymbol{\tau}}^T \rangle_n^{-1} \longrightarrow \sigma^2 \langle \Xi_{\boldsymbol{\tau}}, \Xi_{\boldsymbol{\tau}}^T \rangle^{-1}$ . The consistency of the estimates is implied from the central limit theorem.

**Proof of Proposition 3.3:** To focus on the main point, we ignore the occurrence of false peaks near the boundary because the probability of this event tends to zero as the sample size goes large. We need to prove that  $Prob(MGIC_1(p) < MGIC_1(l)) \longrightarrow 1$  for any integer  $0 \leq l < p$ . We denote the empirical model of order  $p$  fitted from Algorithm 3.1 by  $\hat{y}(t) = \hat{A}_0 + \sum_{j=1}^p \hat{A}_j w_\lambda(t - \hat{\tau}_j)$ . Interestingly, if we replace the observation  $z(t)$  by  $y(t)$  in Algorithm 3.1, then it can be confirmed that the fitted model of order  $p$  is exactly the true model  $y(t) = A_0 + \sum_{j=1}^p A_j w_\lambda(t - \tau_j)$ . Similarly, we denote the theoretical and empirical model of order  $l$  fitted from Algorithm 3.1, using the hypothetical observation  $y(t)$  and real observation  $z(t)$  respectively, by  $y^{(l)}(t) = A_0^{(l)} + \sum_{j=1}^l A_j^{(l)} w_\lambda(t - \tau_j^{(l)})$  and  $\hat{y}^{(l)}(t) = \hat{A}_0^{(l)} + \sum_{j=1}^l \hat{A}_j^{(l)} w_\lambda(t - \hat{\tau}_j^{(l)})$ . We want to prove:  $\|z - \hat{y}\|_n^2 \longrightarrow \sigma^2$  and  $\|z - \hat{y}^{(l)}\|_n^2 \longrightarrow \sigma^2 + c$  in probability, where  $c \geq \inf_{\bar{y}} \|y - \bar{y}\| = d > 0$ , and the infimum is taken over all  $\bar{y} \in SC(w; m; \bar{A}; \bar{\tau})$ , where  $m < p$ . From Theorem 3.3 in LS and Proposition 3.2, we know that  $\{\hat{\tau}_j, j = 1, \dots, p\}$  and  $\{\hat{A}_j, j = 0, \dots, p\}$  are consistent estimates. According to Lemma 4.1,

$$\|\hat{y} - y\|_n \leq \|\hat{A}_0 - A_0\|_n + \sum_{j=1}^p \|\hat{A}_j w_\lambda(t - \hat{\tau}_j) - A_j w_\lambda(t - \tau_j)\|_n \xrightarrow{p} 0.$$

Hence

$$\|z - \hat{y}\|_n^2 = \|\epsilon + y - \hat{y}\|_n^2 = \|\epsilon\|_n^2 + 2\langle \epsilon, y - \hat{y} \rangle_n + \|y - \hat{y}\|_n^2 \xrightarrow{p} \sigma^2,$$

because  $\langle \epsilon, y - \hat{y} \rangle_n \leq \|\epsilon\|_n \|y - \hat{y}\|_n$  and  $\|\epsilon\|_n^2 \xrightarrow{p} \sigma^2$ . Along the same line, we can show that  $\hat{A}_j^{(l)} \xrightarrow{p} A_j^{(l)}$ ,  $\hat{\tau}_j^{(l)} \xrightarrow{p} \tau_j^{(l)}$ , and hence  $\|\hat{y}^{(l)} - y^{(l)}\|_n \xrightarrow{p} 0$ . Theorem 2.1 in LS shows that  $\|y - y^{(l)}\|_n = c \geq d > 0$ . Using the weak law of large numbers for triangular arrays, see Page 35 in Durrett (1991), we can show that  $\langle \epsilon, y - y^{(l)} \rangle_n \xrightarrow{p} 0$ . This implies

$$\|\epsilon + y - y^{(l)}\|_n^2 = \|\epsilon\|_n^2 + 2\langle \epsilon, y - y^{(l)} \rangle_n + \|y - y^{(l)}\|_n^2 \xrightarrow{p} \sigma^2 + c.$$

Putting these together, we have

$$\|z - \hat{y}^{(l)}\|_n^2 = \|\epsilon + y - y^{(l)} + y^{(l)} - \hat{y}^{(l)}\|_n^2$$

$$= \|\epsilon + y - y^{(l)}\|_n^2 + 2 \langle \epsilon + y - y^{(l)}, y^{(l)} - \hat{y}^{(l)} \rangle_n + \|\hat{y}^{(l)} - y^{(l)}\|_n^2 \xrightarrow{p} \sigma^2 + c.$$

Finally

$$\begin{aligned} \text{Prob}[MGIC_1(p) < MGIC_1(l)] &= \text{Prob}[\|z - \hat{y}\|_n^2 + \frac{c_1(n) \log n}{n} p < \|z - \hat{y}^{(l)}\|_n^2 + \frac{c_1(n) \log n}{n} l] \\ &= \text{Prob}[\|z - \hat{y}^{(l)}\|_n^2 - \|z - \hat{y}\|_n^2 > \frac{c_1(n) \log n}{n} (p - l)] \xrightarrow{p} 1. \end{aligned}$$

### Acknowledgment

My work in DNA sequencing and base-calling is carried out jointly with Prof. Terry Speed and Dr. David Nelson. I am especially indebted to Prof. Terry Speed, who has been motivating my research in bioinformatics. Discussions with David Nelson helped my understanding of DNA sequencing. This research is supported by the NSF grant DMS-9971698, and DOE grant DE-FG03-97ER62387. I would like to thank the help provided by the Institute of Pure and Applied Mathematics, UCLA.

### References

- Adams, M. D., Fields, C. and Venter, J. C. editors. (1994). *Automated DNA sequencing and analysis*. Academic Press, London, San Diego.
- An, H. and Gu, L. (1985). On the selection of regression variables. *Acta Math. Appl. Sinica* **2**, 27–36.
- Berno, A. J. (1996). A graph theoretic approach to the analysis of DNA sequencing data. *Genome Research* **6**, 80–91.
- Berno, A. J. and Stein, L. (1995) *Bass manual*. Stanford University.

- Cawley, S. E. (2000). *Statistical models for DNA sequencing and analysis*. PhD thesis, University of California, Berkeley.
- Chen, W.-Q. and Hunkapiller, T. (1992). Sequence accuracy of larger DNA sequencing projects. *J. DNA Sequencing and Mapping* **2**, 335–342.
- Durrett, R. (1991). *Probability: Theory and examples*. Wadsworth & Brooks/Cold, Pacific Grove, California.
- Ewing, B. and Green, P. (1998). Base-calling of automated sequencer traces using *phred*. 2. error probabilities. *Genome Research* **8**, 186-194.
- Ewing, B., Hillier, L., Wendl, M. C. and Green, P. (1998). Base-calling of automated sequencer traces using *phred*. 1. accuracy assessment. *Genome Research* **8**, 175-185.
- Giddings, J. C. (1965). *Dynamics of chromatography*. Marcel Dekker, New York.
- Giddings, M. C., Brumley, R. L., Haker, M. and Smith, L. M. (1993). An adaptive, object oriented strategy for base calling in DNA sequence analysis. *Nucleic Acids Research* **21(19)**, 4530–4540.
- Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations*. The John Hopkins University Press: Baltimore and London, 3rd edition edition.
- Grossman, P. D., Menchen, S. and Hershey, D. (1992). Quantitative analysis of DNA sequencing electrophoresis. *Genetic Analysis, Techniques, and Applications* **9**, 9-16.
- Huang, W., Yin, Z., Fuhrmann, D. R., States, D. J. and Thomas, L. J. (1997). A method to determine the filter matrix in four-dye fluorescence-based DNA sequencing. *Electrophoresis* **18**, 23–25.

- Ives, J. T., Gesteland, R. F. and Stockham, T. G. (1994). An automated film reader for DNA sequencing based on homomorphic deconvolution. *IEEE Transactions on Biomedical Engineering* **41(6)**, 509–519.
- Kheterpal, I., Li, L., Speed, T. P. and Mathies, R. A. (1999). A three-color labeling approach for DNA sequencing using energy transfer primers and capillary electrophoresis. *Electrophoresis* **19**, 1403–1414.
- Koop, B. F., Rowen, L., Chen, W.-Q., Deshpande, P., Lee, H. and Hood, L. (1993). Sequence length and error analysis of sequence and automated *taq* cycle sequencing methods. *BioTechniques* **14(3)**, 442–447.
- Lander, E. S. and Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones. *Genomics* **2**, 231–239.
- Li, L. (1998). *Statistical Models of DNA Base-calling*. PhD dissertation, University of California, Berkeley.
- Li, L. and Speed, T. P. (1999). An estimate of the color separation matrix in four-dye fluorescence-based DNA sequencing. *Electrophoresis* **20**, 1433–1442.
- Li, L. and Speed, T. P. (2000). Parametric deconvolution of positive spike trains. *Annals of Statistics*, in press.
- Li, L. and Speed, T. P. (2001). Deconvolution of sparse positive spikes: is it ill-posed? Technical report, Department of Statistics, University of California, Berkeley.
- Lawrence, C. B. and Solovyev, V. V. (1994). Assignment of position-specific error probability to primary DNA sequence data. *Nucleic Acid Research* **22(7)**, 1272–1280.

- Luckey, J. A., Norris, T. B. and Smith, L. M. (1993). Analysis of Resolution in DNA sequencing by capillary gel electrophoresis. *Journal of Physical Chemistry* **97**, 3067-3075.
- Lumpkin, O. J., DeJardin, P. and Zimm, B. H. (1985). Theory of gel electrophoresis of DNA. *Biopolymers* **24**, 1573–1593.
- Nelson, D. O. (1996). Improving DNA sequence accuracy and throughput. In T. P. Speed and M. S. Waterman, editors, *Genetic Mapping and DNA Sequencing*, volume 81 of *The IMA Volumes in Mathematics and its Applications*, pages 183–206. Springer.
- Nelson, D. O. and Speed, T. P. (1996). Recovering DNA sequences from electrophoresis. In Levinson, S. E. and Shepp, L., editors, *Image Models (and their Speech Model Cousins)*, pages 141–152. Springer-Verlag.
- PE Applied Biosystems Inc., Foster City, CA. (1996). *ABI PRISM, DNA sequencing analysis software*.
- Riesz, F. and Nagy, B. Sz. (1955). *Functional Analysis*. Ungar, New York.
- Russell, P. J. (1995). *Genetics*. Harpercollins College Publisher, New York.
- Tibbetts, C., Bowling, J. M. and Golden, J. B. (1994). Neural networks for automated base-calling of gel-based DNA sequencing ladders. In *Automated DNA sequencing and analysis*, pages 219–230. Academic Press: London, San Diego.
- Waterman, M. S. (1995). *Introduction to Computational Biology: Maps, Sequences and Genomes*. Chapman & Hall, London

Yin, Z., Severin, J., Giddings, M. C., Huang, W. and Westphall, M. S. and Smith, L. M. (1996). Automatic matrix determination in four dye fluorescence-based sequencing. *Electrophoresis* **17**, 1143–1150.

Yu, B. and Speed, T. P. (1997). Information and the clone mapping of chromosomes. *Annals of Statistics* **25**, 169–185.

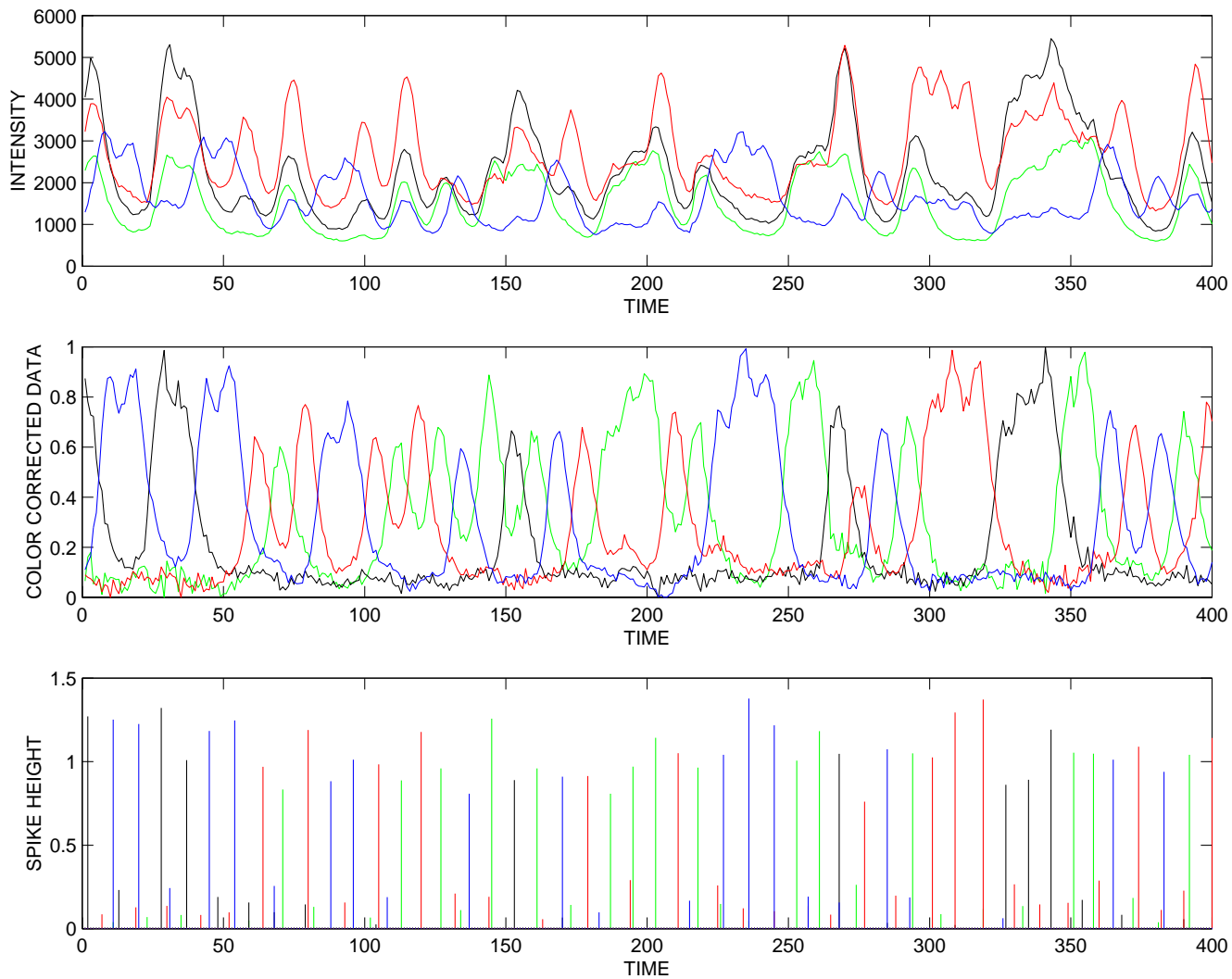


Figure 1: Top: a segment of real DNA sequencing data — fluorescence intensities; Middle: the color-corrected data — dye concentrations, with proper normalization; Bottom: the output from parametric deconvolution — a Dirac delta train, representing the occurrences of the nucleotide bases.

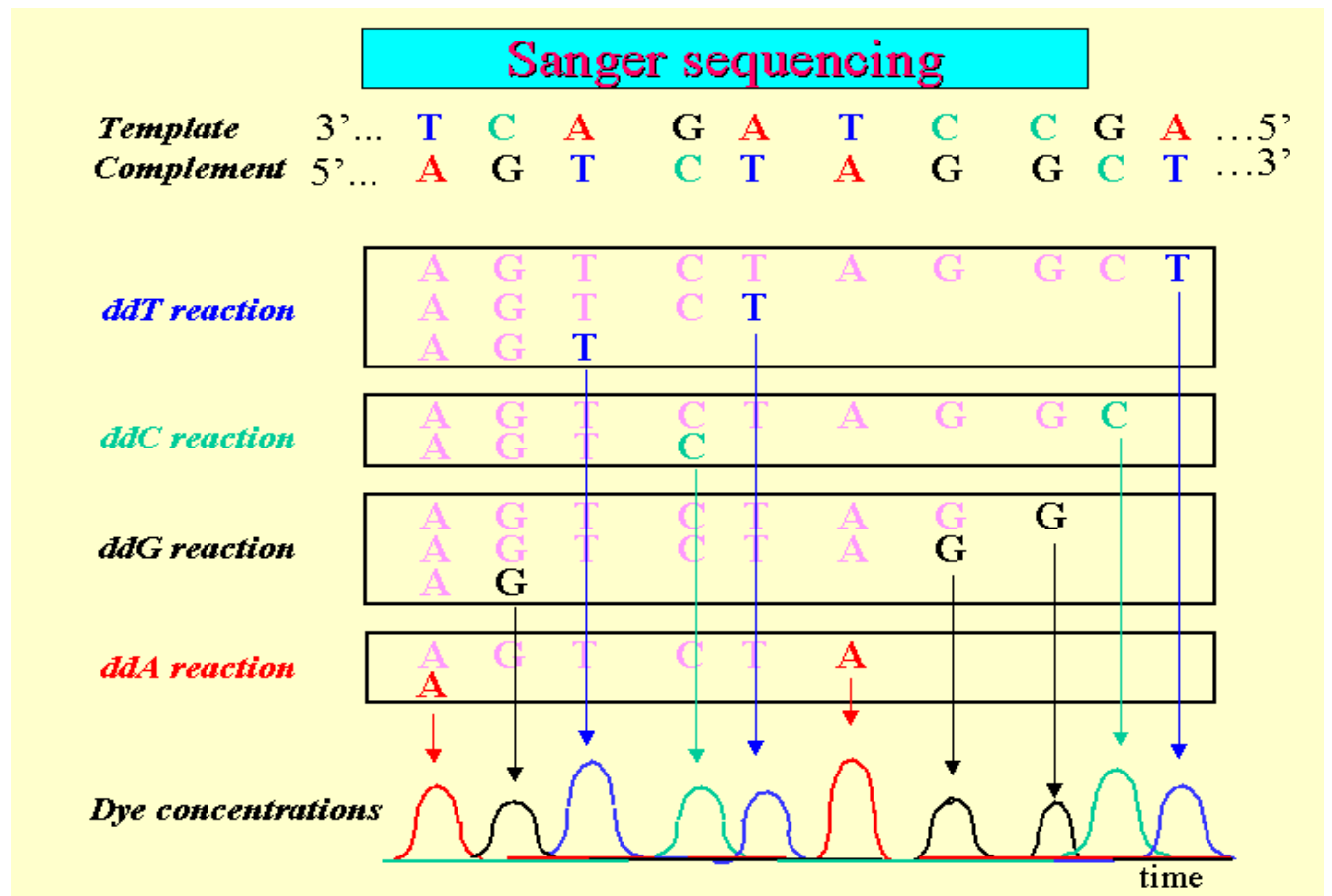


Figure 2: A schematic representation of Sanger sequencing.

## A modeling framework of DNA sequencing and base-calling

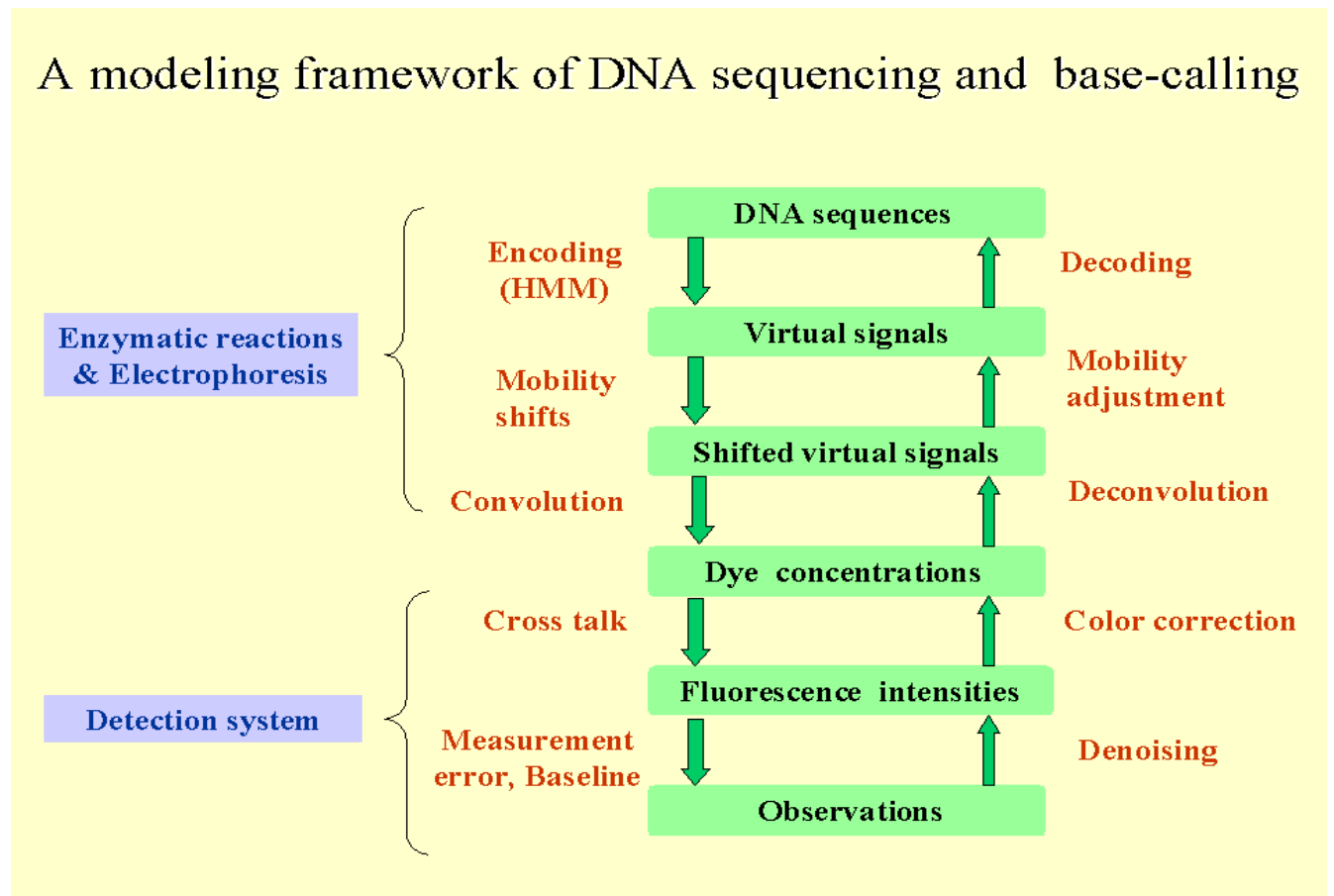


Figure 3: A modeling framework of DNA sequencing and base-calling.

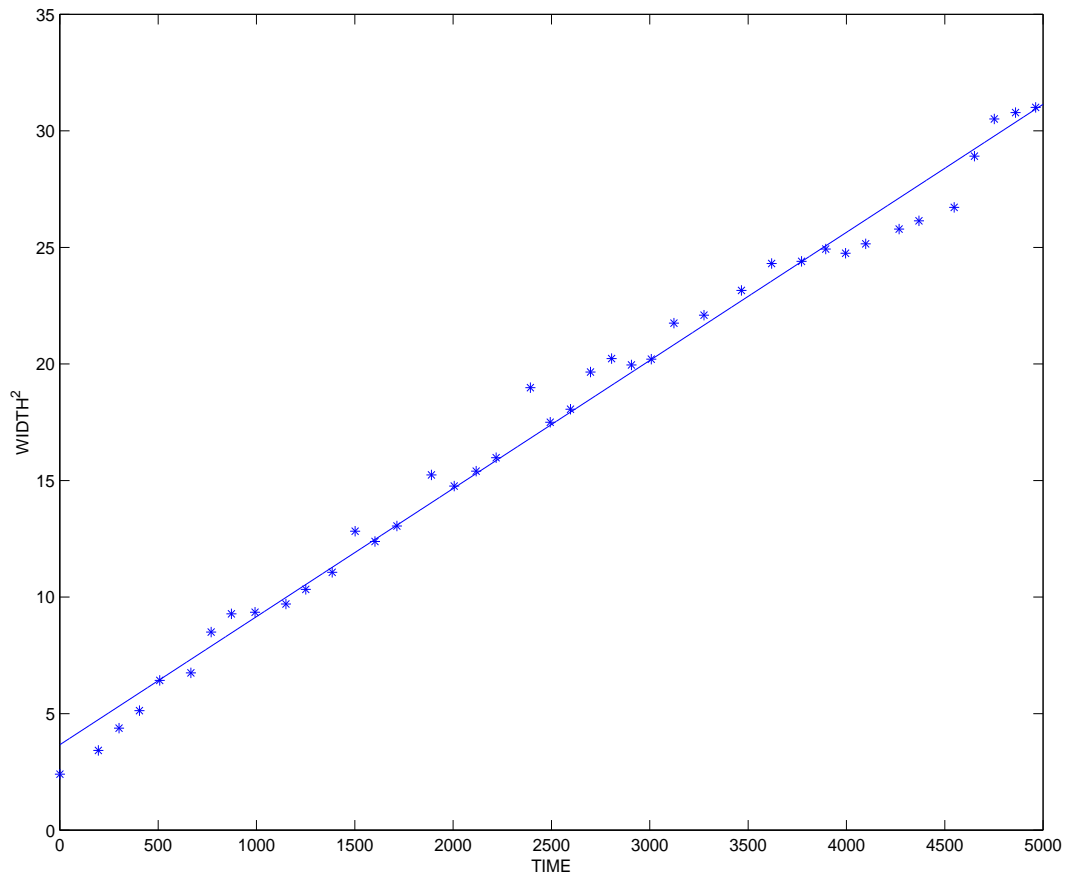


Figure 4: The square of estimated width parameter versus time.