

ITERATIVE DECONVOLUTION FOR AUTOMATIC BASECALLING OF THE DNA ELECTROPHORESIS TIME SERIES

Xiao-Ping Zhang[†], Daniel Allison

Applera Corporation -- Applied Biosystems
850 Lincoln Centre Drive, Foster City, CA 94404, USA

ABSTRACT

In DNA (deoxyribonucleic acid) sequencing, there are four possible chemical base types: adenine (A), cytosine (C), guanine (G), thymine (T), which contain genetic information. The four base types are identified by examining four DNA electrophoresis time series. This procedure is called “basecalling”. However, in practice, there are many other undesired signal features that prevent the accurate base detection in the electrophoresis time series. In this paper, a systematic signal processing method is developed in enhancing the signal quality of the DNA electrophoresis time series based on an iterative deconvolution method. It recovers the sharp peaks in all signal time series and subsequently improves the basecalling performance.

1. INTRODUCTION

In DNA (deoxyribonucleic acid) sequencing applications, there are four possible chemical base types: adenine (A), cytosine (C), guanine (G), thymine (T), which contain genetic information. In chemical processing of DNA sequences, electrophoresis is used to discriminate the molecules by length, which can be translated to the position of each base [1]. Each base in the obtained DNA electrophoresis time series is represented by high-level signals

(peaks) with certain shape. By looking for the positions of these signal “peaks” in the four DNA electrophoresis time series, the DNA base sequence is therefore identified. This procedure is called “basecalling”. Ideally, at any base position, there should be a peak in only the corresponding electrophoresis time series. However, in practice, there are many other undesired signal features that prevent the accurate peak detection in the electrophoresis time series. The most prominent fact is the degradation of signal resolution, i.e., the signal peak is not an ideal sharp peak but is a waveform with certain spread width. When there are multiple consecutive peaks, it becomes difficult to correctly detect the accurate signal peaks. This problem becomes severe close to end of the time series since the signal resolution becomes very poor.

In this paper, a new practical systematic signal processing method in enhancing the signal quality of the DNA electrophoresis time series is reported. It recovers the sharp peaks in all signal time series and subsequently improves the basecalling accuracy and read length. The core of the method is an iterative nonlinear deconvolution algorithm. Unlike the commonly used linear Wiener filtering method, which often suffers ringing effects, the new method recovers the sharp base peaks without adding any secondary false peaks caused by ringing. The

[†] Contact for X.-P. Zhang: Department of Electrical and Computer Engineering, Ryerson University, 350 Victoria Street, Toronto, Ontario, Canada, M5B 2K3. Email: xpzhang@ieee.org

following sections of the paper will briefly describe the basic concepts, formulation, and principles of this method.

2. FORMULATION OF THE DECONVOLUTION PROBLEM IN DNA ELECTROPHORESIS TIME SERIES

Ideally, the electrophoresis signal generation for the DNA sequencing can be treated as a linear system. It is assumed that the observed DNA electrophoresis digital signal $y(n)$ is the convolution of the input signal $x(n)$ and point spread function $h(n)$

$$y(n) = x(n) \otimes h(n). \quad (1)$$

In electrophoresis, $x(n)$ is supposed to be a sparse pulse train which represents the base locations and signal stimulate amplitude, i.e.,

$$x(n) = \sum_k a(k)p(n-k), \quad (2)$$

where the $p(n)$ is a very narrow pulse, in ideal case $p(n)=\delta(n)$, where $\delta(n)$ is Kronecker function, $a(k)\neq 0$, and k represents the base positions. The model can be shown in Figure 1. The basecalling objective is then to find an accurate estimate of $x(n)$, denoted by $\hat{x}(n)$, given the observed electrophoresis series $y(n)$.

3. ITERATIVE DECONVOLUTION ALGORITHM FOR DNA ELECTROPHORESIS TIME SERIES

3.1 Iterative algorithm

An iterative algorithm can be generally described as follows [2]. Starting from an initial signal vector \mathbf{x}_0 and using the following iteration,

$$\mathbf{x}_{k+1} = F\mathbf{x}_k, \quad (3)$$

where “ F ” is an operator and \mathbf{x}_k denotes the signal vector value at the k -th iteration, we expect to find an operator such that when k is sufficiently large, \mathbf{x}_k converges to the underlying pulse train represented by Eq. (2), denoted by vector \mathbf{x} , i.e.,

$$\lim_{k \rightarrow \infty} \mathbf{x}_k = \mathbf{x} \quad (4)$$

In the iterative algorithm as in Eq. (3), if “ F ” is a contract mapping with \mathbf{x} being the fixed point of the mapping, i.e.,

$$\|F\mathbf{x}_i - F\mathbf{x}_j\| \leq r\|\mathbf{x}_i - \mathbf{x}_j\|, 0 \leq r < 1 \quad (5)$$

and

$$\mathbf{x} = F\mathbf{x}, \quad (6)$$

the iterative algorithm will converge to \mathbf{x} , i.e.,

$$\lim_{k \rightarrow \infty} \mathbf{x}_k = \mathbf{x}.$$

A basic iteration equation is [2]:

$$\mathbf{x}_{k+1} = F\mathbf{x}_k = \lambda\mathbf{y} + G\mathbf{x}_k, \quad (7)$$

where operator G should be constructed according to the applications such that F is a contract mapping.

3.2 Design of the contract mapping for electrophoresis time series

We explore some major properties of the DNA electrophoresis time series, such as the following constraints:

- Positivity – We suppose that the underlying pulse train which represents the bases are always positive, i.e., $a(k) \geq 0$ in Eq. (2).
- Time localization – We suppose that each pulse $p(n)$ that represents one base has limited time duration (i.e., pulse $p(n)$ must be very narrow). We denote the duration of one pulse as d . Note that in extreme case – when $p(n)=\delta(n)$, we have $d=1$.

By incorporating such specific signal properties of the DNA electrophoresis, a contracted mapping operator G is developed for the DNA electrophoresis signal vector \mathbf{x} (with individual element denoted by $x(n)$). Then a mapping F can be constructed:

$$\mathbf{x}_{k+1} = F\mathbf{x}_k = G\mathbf{x}_k + \lambda(\mathbf{y} - h \otimes G\mathbf{x}_k) \quad (8)$$

It is easy to verify that \mathbf{x} is a fixed point of F , if \mathbf{x} is a fixed point of mapping operator G . It can also be proved if the point spread function h satisfies certain property, the mapping F is a contract mapping [3]. We will not elaborate the details in this summary for the limit space.

Note that the operators we defined are nonlinear. Therefore, the constructed iterative algorithm is a nonlinear method and not a linear filtering method.

4. MAJOR ALGORITHMIC PROCEDURES

The following outline of the general procedures is used to obtain the deconvolved DNA electrophoresis signals.

- 1) *Preprocessing of the electrophoresis signal:* The four dye electrophoresis signals are filtered and multicomponented and the baseline is removed. The mobility shift is compensated and the peak spacing is normalized along the time dimension. This produces a regularized signal that should be similar for all input samples.
- 2) *Adaptive point spread function estimation:* Peaks are detected in the regularized trace and called as bases with standard classification methods. The called peaks are used to adaptively estimate the local point spread function h . The time-localization parameter d is estimated according to the peak spacing in the segment.
- 3) *Adaptive deconvolution:* The iterative deconvolution algorithm developed in section 3 is applied adaptively according to the estimated local point spread function.
- 4) *Output:* The deconvolved signal array is output and used for final basecalling.

5. EXAMPLES

Figure 2 shows an illustration of the developed new algorithm applying on a piece of electrophoresis signal. Figure 3 illustrates a real electrophoresis sequencing sample file (generated by 3100 Genetic Analyzer) processed by the basecalling algorithm with the developed new iterative deconvolution algorithm incorporated. In Figure 3, the original processed data without the new method is shown on top and the data processed with this new method is

shown on the bottom. The improvement of the signal quality is evident in these examples.

6. CONCLUSIONS

We developed a new systematic signal processing method to enhance the quality of the raw DNA sequencing electrophoresis signals. The enhanced signal may be used to improve the performance of the basecalling of the DNA sequence and other sequence analysis applications. The core algorithm of this new method is a nonlinear iterative deconvolution algorithm incorporating the specific characteristics. This algorithm proved to be very effective to recover sharp peaks at base positions in DNA sequencing electrophoresis signals without introducing extra small peaks (ringing) as in conventional linear Wiener filter based signal filtering methods. Adding this deconvolution into the processing of an existing basecaller reduces the total basecalling error rate by 1% - 2%. It is noted that significant basecalling accuracy improvements are shown in low resolution signal areas. Further modifications to the basecaller to take full advantage of the narrow deconvolved peaks may further reduce this error rate and significantly improve the overall accuracy and read length.

7. REFERENCES

- [1] T. A. Brown, *DNA Sequencing: The Basics*, Oxford University Press, New York, 1994.
- [2] R. W. Schafer, R.M. Mersereau and M. A. Richards, "Constrained Iterative Restoration Algorithms," *Proceedings of the IEEE*, vol. 69, no. 4, April 1981.
- [3] X.-P. Zhang, D. B. Allison, "Iterative deconvolution for automatic basecalling in sequencing analysis," *Technical Report*, Applied Biosystems, Apr. 2002.

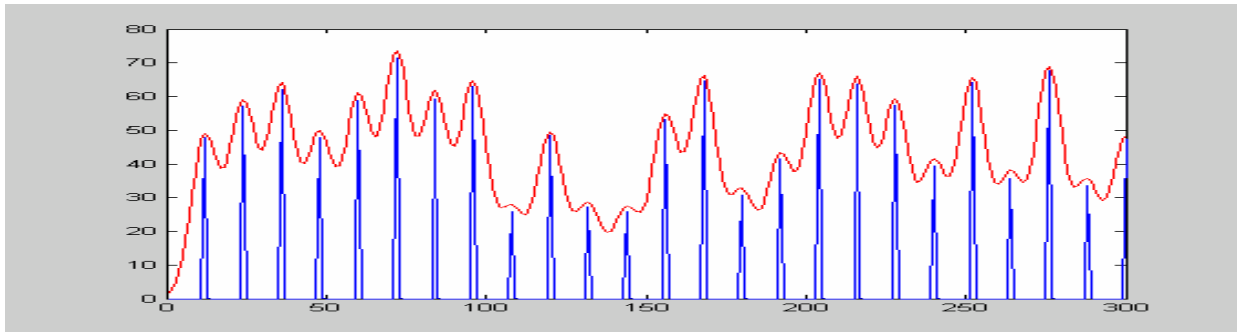


Figure 1. The pulse train (blue line) represents $x(n)$, and the smooth line (red line) represent the observed DNA electrophoresis signal $y(n)$

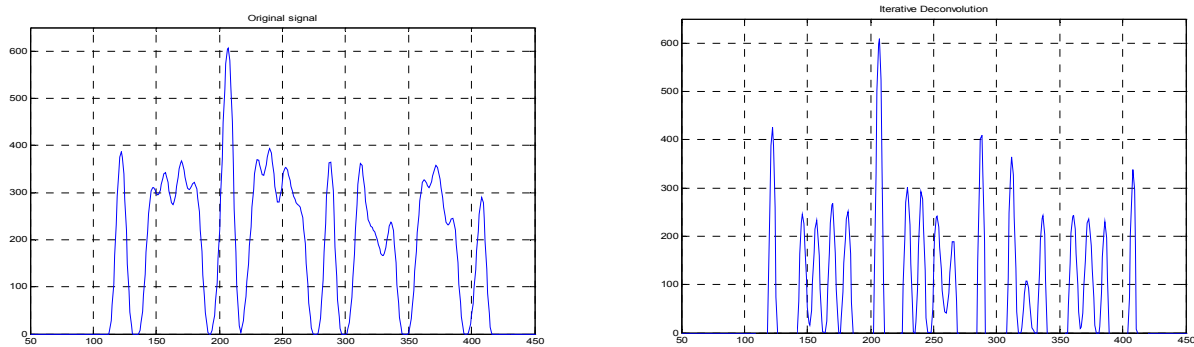


Figure 2. (a) Original signal; (b) Deconvolved signal by the proposed algorithm

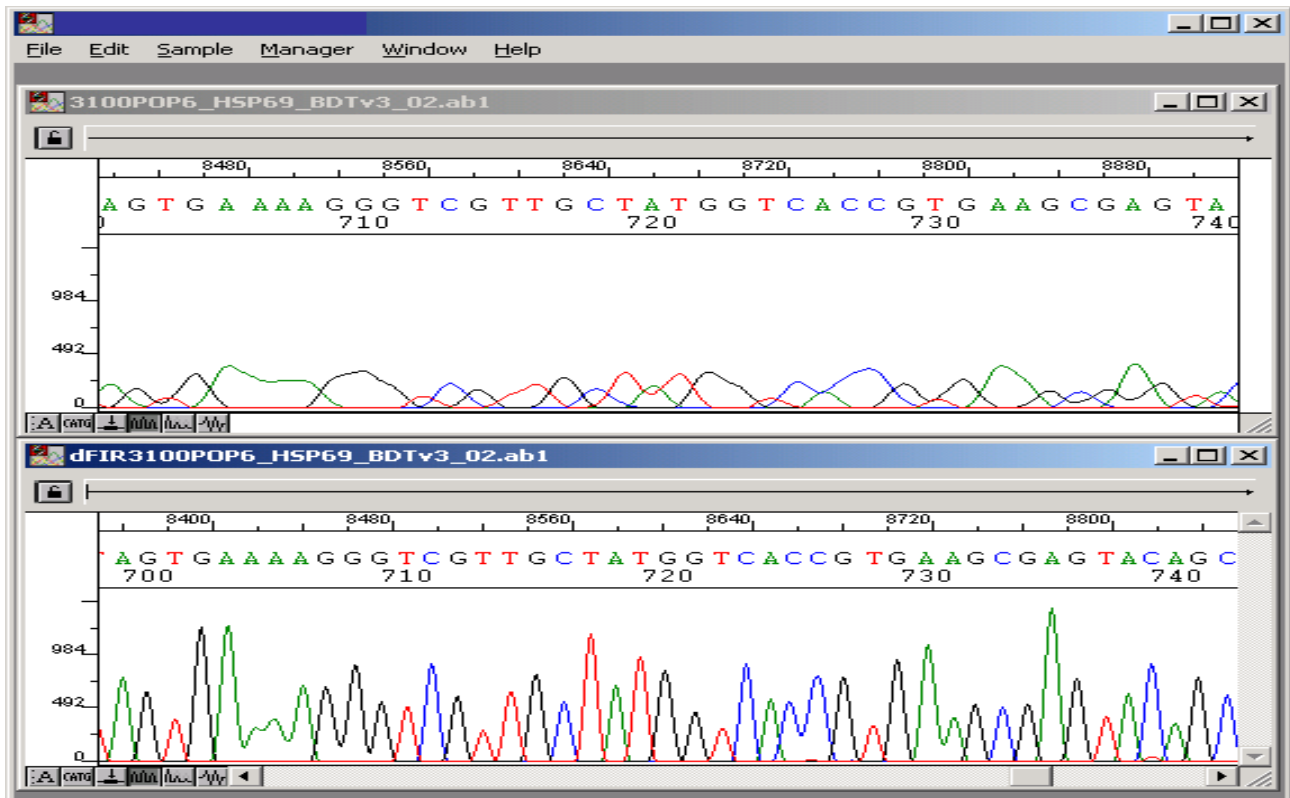


Figure 3. Deconvolved data (bottom one) for a real sequencing file generated by 3100 Genetic Analyzer (top one is the original data.)