12-11-2000

An Intelligent System for Automated DNA Base Calling

Mohamad T. Musavi
Principal Investigator; University of Maine, Orono, musavi@maine.edu

Final Report for Period: 07/1999 - 06/2001 **Submitted on:** 12/11/2000

Principal Investigator: Musavi, Mohamad T. **Award ID:** 9902565

Organization: University of Maine

An Intelligent System for Automated DNA Base Calling

Project Participants

Senior Personnel

Name: Musavi, Mohamad
Worked for more than 160 Hours: Yes
Contribution to Project:

Post-doc

Name: Domnisoru, Cristian
Worked for more than 160 Hours: Yes
Contribution to Project:
Postdoctoral research in DNA base calling software.
Support from NSF-NATO.

Graduate Student

Name: Zhan, Xiang
Worked for more than 160 Hours: Yes
Contribution to Project:
The person's M.S. thesis involved test and evaluation of the techniques developed in this research. The person was supported through another existing grant.

Undergraduate Student

Organizational Partners

DNA Sequencing Core Facility at U. Maine
Thanks to the DNA Sequencing Core Facility at University of Maine, we get a steady source of DNA sequencing data produced by the ABI Prism 373 machine.

Other Collaborators or Contacts

Dr. Jurgan Naggert and Mr. Douglas McMinimy of The Jackson Laboratory.

Activities and Findings

Project Activities and Findings:
A novel algorithm for base calling has been developed and tested on a limited database. The algorithm is based on 1) a novel data preprocessing to ensure that the initial information contained in the data sets is preserved through filtering, and 2) a prediction of the position of the following base at each step. For predicting the base spacing, a model is constructed from the good region of the trace.

Project Training and Development:
- A 'mechanical shift' in DNA data collection has been identified. Appropriate steps for compensating this shift have been devised.

- A set of filtering steps has been proposed to eliminate the cross-talk between the four signals A, C, G, and T in the trace.

- A specific pattern in base spacing has been identified. For each trace file, a model of the base spacing is detected allowing for a prediction of the space between bases at the end of the sequence where the data is not as good as in the begining.

Research Training:

During this project Dr. Cristian Domnisoru and the graduate student involved had the opportunity to:
1) not only develop research capabilities in bioinformatics but also become familiar with the DNA sequencing process and base calling.
2) teach a course in the University of Maine Electrical & Computer Engineering Department (ECE 323), thus enhancing his teaching experience.

Outreach Activities:

Journal Publications

Cristian Domnisoru, Xiang Zhan, Mohamad Musavi, "Cross-talk Filtering in Four Dye Fluorescence-based DNA Sequencing", *Electrophoresis*, p. 2983, vol. 21/14, (2000).) Published
Cristian Domnisoru, Mohamad Musavi, "Base Calling Algorithm for DNA Electrophoresis Sequencing", *Electrophoresis*, p. , vol. , ().) Submitted

Books or Other One-time Publications

C. Domnisoru and M.T., Musavi, "Filtering Technique for Fluorescence-Based DNA Sequencing Data", (2000). *Conference*, Published
Collection: Proceedings of IEEE-CCECE 2000 Canadian Conference on Electrical and Computer Engineering, Halifax, Nova Scotia, Canada
Bibliography: May 7-10, 2000, pp.489-493.

C. Domnisoru and M.T. Musavi, "Mechanical Shift and Base Spacing Modelling and Compensation for DNA Sequencing Raw Data Processing", (2000). *Conference*, Published
Collection: Proceedings of the MS'2000 Modelling and Simulation, IASTED International Conference
Bibliography: May 15-17, 2000 Pittsburgh, Pennsylvania, USA, pp.470-476.

C. Domnisoru and M.T., Musavi, "Mechanical Shift Compensation for DNA Trace File Processing", (2000). *Poster*, Published
Collection: Poster presentation at the third International Symposium on Capillary Electrophoresis and Related Microscale Techniques, Hong Kong
Bibliography: June 14-17, 2000.

C. Domnisoru and M.T., Musavi, "Base Calling Algorithm for DNA Electrophoresis Sequencing", (2000). *Plenary presentation*, Published
Collection: Plenary presentation at the third International Symposium on Capillary Electrophoresis and Related Microscale Techniques, Hong Kong
Bibliography: June 14-17, 2000

C. Domnisoru and M. Musavi, "Adaptive Base Calling Strategy", (2000). *Poster*, Published
Collection: Poster presentation at the third International Symposium on Capillary Electrophoresis and Related Microscale Techniques, Hong Kong
Bibliography: June 14-17, 2000

C. Domnisoru and M.T. Musavi, "Mobility Shift Compensation in DNA Base Calling", (2000). *Poster*, Published
Collection: Poster presentation at the German Conference in Bioinformatics, Heidelberg, Germany
Bibliography: October 5-7, 2000

C. Domnisoru and M. Musavi, "Basecalling Algorithm for DNA Electrophoresis Sequencing Data", (2000). *Plenary presentation*, Accepted
Collection: Fourth Annual Conference on Computational Genomics
Bibliography: November 16-19, 2000

Xiang Zhan, "DNA Basecalling using Neural Networks", (2000). *Thesis*, Published
Collection: M.S. Thesis
Bibliography: University of Maine Library

Web/Internet Sites

URL(s):
http://www.eece.maine.edu/Research/IntSys/dna.htm
Description:
Brief description of the project and publications.

Other Specific Products

Product Type: Data or databases
Product Description:
Five hundred DNA trace files including their correct sequence have been collected.

Sharing Information:
The data will be provided to others on request.

Product Type: Software (or netware)
Product Description:
Base calling software modules have been developed in C language.

Sharing Information:
The software will be made available to others on request.

Contributions

Contributions within Discipline:
An investigation into improving the performance of DNA base calling algorithms was conducted. The results have shown that the preprocessing steps performed by ABI sequencer on raw data adversely affects the accuracy of DNA sequencing. This adverse effect has been responsible for relatively high error rates, between 3.5% to 6%, in both ABI and Phred sequencing software. Please note that Phred also uses the processed data generated by ABI sequencer; only their base-calling algorithm is different. To remedy this effect, we have developed and implemented a new filtering technique that preserves the initial information contained in the raw data. This provides qualitatively superior data for the future base calling step. Our proposed filtering step provides mechanical shift compensation, cross-talk filtering, and baseline adjustment. These have been briefly described below. Application of our filtering step on a limited number of DNA data has provided sequences with lower error rate.
Contributions to Other Disciplines:
From one of our presentations in Heidelberg, Germany, we found out that the methodologies developed for DNA base calling would also be beneficial to crystallography research.
Contributions to Human Resource Development:
In addition to the NATO visiting scientist, one graduate and two undergraduate students were also involved in the project. The research has so far resulted in one journal and five reviewed conference papers and one Master of Science thesis. The research also resulted in software that has been used in conducting the preliminary investigation of the current proposal. The collaboration between the engineering research team and the University of Maine Core DNA sequencing facility provided an opportunity for the team to also become familiar with the biological aspects of DNA data preparation and processing.
Contributions to Science and Technology Infrastructure:
This research project prepared us to ivestigate other bioinformatic areas and collaborate with several researchers from different disciplines. As the result, we are now investigators in a multidisciplinary NIH IDEA proposal intending to develop intelligent information processing techniques for handling gene expression data.
Beyond Science and Engineering:
We have a pending NSF proposal to complete our current work and develop a complete base calling software. The result of this pending proposal is suitable for commercislization.

Categories for which nothing is reported:

Activities and Findings: Any Outreach Activities

Project Report

a. NSF award number: DGE-9902565

The duration of this project was one year from July 1999 to June 2000 for a total of $46,200.

b. Title of the project

An Intelligent System for Automated DNA Base Calling

c. Summary of the results:

An investigation into improving the performance of DNA base calling algorithms was conducted. The results have shown that the preprocessing steps performed by ABI sequencer on raw data adversely affects the accuracy of DNA sequencing. This adverse effect has been responsible for relatively high error rates, between 3.5% to 6%, in both ABI and Phred sequencing software. Please note that Phred also uses the processed data generated by ABI sequencer; only their base-calling algorithm is different. To remedy this effect, we have developed and implemented a new filtering technique that preserves the initial information contained in the raw data. This provides qualitatively superior data for the future base calling step. Our proposed filtering step provides mechanical shift compensation, cross-talk filtering, and baseline adjustment. These have been briefly described below. Application of our filtering step on a limited number of DNA data has provided sequences with lower error rate.

i) *Mechanical shift*

The ABI Prism 373 DNA Sequencer machine uses four dyes that fluoresce at different wavelengths. Since there is an overlap in the emission spectra, the four data sets will reflect a cross-influence or cross-talk. To compensate for this overlap, ABI has built a filtering process in their software. However, ABI's filtering process as well as the one presented by Giddings et al is affected by noise. This noise shows up as a small (false) signal in places where there already exists a dominant peak. This noise shows its adverse effect especially at the end of the sequence data where the amplitude of the real data is small. Also, the noise appears to be significantly stronger at the bottom of the signal while not affecting the peaks. Both ABI and Giddings et al try to eliminate this noise by filtering-out the high frequencies using a Fourier based technique to cut out the bottom part of the signal.

We consider two major drawbacks with their approach. First, eliminating the high frequencies will also attenuate the separation between two consecutive peaks of the same base, especially near the primer. Second, cutting-out a part of the signal will affect the shape of the peaks, especially at the end of the sequence (the signal with a Gaussian-like distribution turn into a dome-like one). We consider that these data manipulation techniques have a negative influence on the accuracy of the final data used for the base call. In other words, the base caller module will have less significant data than the originally scanned one.

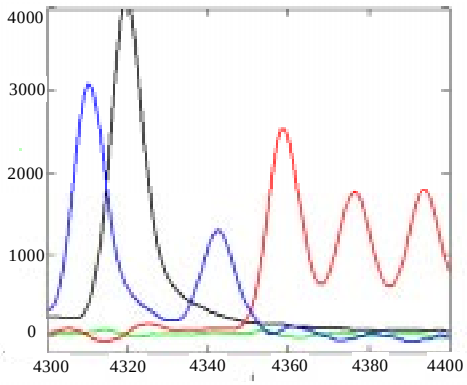


Fig.1: Before mechanical shift

Another important phenomenon noticeable in the raw data is that the bottom of the signal does not maintain a constant value. It changes along the sequence in a limited range making it difficult to apply a matrix for cross-influence filtering. Consequently, ABI software and Giddings et al are applying another filter, this time to cut out the low frequencies. We consider that this approach modifies the data, changing the shape of the peaks for the cases where several consecutive peaks of a base are at the end of the sequence. This case is treated like a low frequency component and is discarded when, in fact, the data is just good as it is.

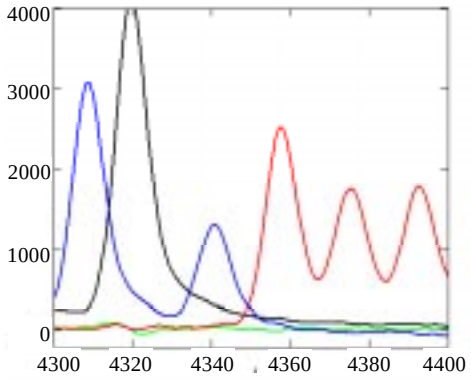


Fig.2: After mechanical shift compensation

In our approach, we have based our signal processing on the variation of the signal from one point to the other and not on the instantaneous value. We have built a matrix of coefficients reflecting the mutual influence between the 4x4 possible pairs of bases. We have applied the matrix for cross-influence removal to the variation of the signal and then recomposing it back, thus, avoiding the use of a low frequency filter. However, the previously mentioned noise at the bottom of the signal is still present as it can be seen in Fig. 1.

The main problem is that the signal corresponding to one base contains small peaks when another dominant signal is present. For example, Fig. 1 shows that the signal for base **C** (blue) after the 2nd **C** base still contains some peaks even when a dominant **T** (red) base is present. Also note that the frequency of small **C** peaks are similar to that of the signal **T**. The two signals, **C** and **T**, "appear" to be shifted. Consequently, since the cross-influence matrix is a constant one, we cannot completely remove the mutual influence between the signals. We found out that by "shifting" the four signals with appropriate intervals, the noise can be significantly reduced. Fig. 2 presents the same region as in Fig. 1 except that the data was shifted in order to obtain a minimum quantity of noise at the bottom of the signal. It should be noted that the same cross-influence matrix has been applied to both figures to show the effectiveness of the shifting step.

ii) *Cross-talk filtering*

Since there is an overlap in the emission spectra, the four data sets will reflect a cross-influence, which is called "Cross-talk". To get rid of this effect, a multi-component analysis, or deconvolution process, must be added before we make base calls. To deconvolve the measured signals vector ***s*** and obtain the four actual fluorescence intensities vector ***f***, which are in fact the measures of dye concentration in the detection region, Eq. (1) has been used by reseachers:

$$\mathbf{f} = \mathbf{M}^{-1} \cdot \mathbf{s} \qquad (1)$$

M is a 4x4 matrix that needs to be determined. The most promising approaches for finding **M** are based on the information contained in the raw data itself. These techniques require that the signals contained in vector ***s*** be in a correct relative position with respect to each other and the baseline. This necessitates a baseline adjustment before cross-talk removal. This adjustment is due to the fact that the four signals are instantaneous values. Unfortunately, the recorded information is not aligned. Various factors account for a drift of each of the signals with different quantities. First, the fluorescence signals collected have different baseline levels for the four different wavelengths. Also, the deformation of the gel as a result of temperature changes,

and the variation of the laser output power, can determine a slow varying baseline of the four signals.

In our studies, we have found that the main problem related to the baseline adjustment is the difficulty to separate three main sources of baseline variation. The first is the change in the background lighting during data collection. The second is the compound effect of occurrence of consecutive bases of the same type, and the third is due to cross-talk between different signals. Since we cannot separate between the three sources, the requirement to align the signal to the baseline might be contributing to a distortion of the signal. In fact, we have found out that the raw data itself does not need to be perfectly aligned in order to have a perfect deconvolution.

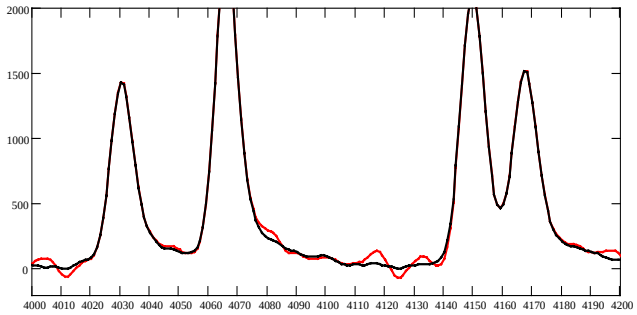


Fig.3: Base T before (red) and after (black) the second filtering step

To address the baseline adjustment dilemma, we have modified the algorithm traditionally employed by two filtering steps. These modifications are: 1) use of the signal variation in the deconvolution process instead of the actual value, 2) changing the order of performing baseline adjustment and cross-talk removal, and 3) accounting for the nonlinearity of the transformation between the measured signal and the actual signal. In the first filtering step, we use a deconvolution matrix of coefficients M.

Our second filtering step is based on the observation that the transformation from the "detector space" or raw data to the "filtered space" is non-linear. An approach similar to the general multi-component analysis, Eq. (1), was employed, as given by Eq. (2.

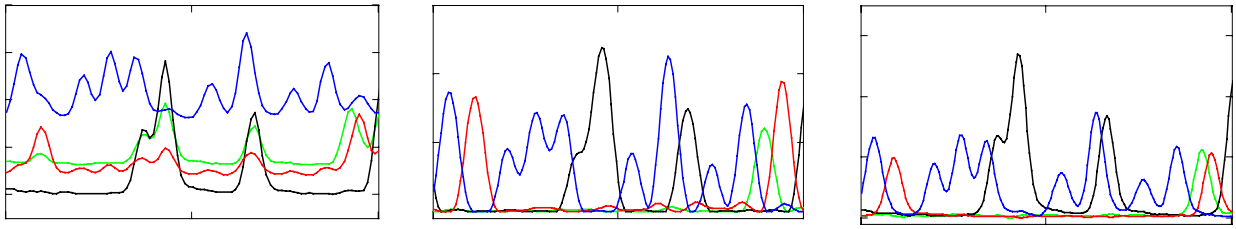
$$\underline{\underline{\Delta \mathbf{s}}} = \underline{\Delta \mathbf{s}} + \mathbf{T} \cdot \underline{\Delta \mathbf{s}'} \qquad (2)$$

In Eq. (2), $\underline{\Delta s}$ and $\underline{\Delta s}'$ are the signals and their derivatives respectively after the deconvolution with matrix M. The signal $\underline{\underline{\Delta \mathbf{s}}}$ is the result of the second filtering step. The end result of this step, the signals $\underline{\underline{\Delta g}}$, $\underline{\underline{\Delta c}}$, $\underline{\underline{\Delta t}}$ and $\underline{\underline{\Delta a}}$ are then used to reconstruct back the signals of the four fluorophores.

Our research resulted in an automatic method for obtaining matrices **M** and **T** described above. In Fig.3 we present a small fragment of the signal for base **T** before (red) and after (black) the second filtering step. Attenuation of the false peaks in the signal as the result of our filtering step is significant.

To compare the overall result of our filtering procedure, mechanical shift compensation, cross-talk removal, and baseline adjustment with the ABI result, we present a section of DNA raw data in Fig. 4 (A). The ABI's filtered data is presented in Fig. 4 (B) while ours is given in Fig. 4 (C). The first observation is that the ABI software provides signals with noise at the baseline. This can be noticed by examining signal **T** (red). The second observation refers to the two consecutive **G** (black) peaks in the center of Fig. 4 (B) and (C). While in the ABI's result the two peaks are not clearly separated, our result shows two well defined peaks. This becomes extremely important in the subsequent base calling modules. It is clear that providing the base

calling modules with improved preprocessed data, containing well-defined peaks, the errors of the overall process can be significantly reduced.



(A) (B) (C)

Fig.4: Comparing Raw data (A), ABI's processed data (B) and our processed data (C)

d. Contribution to the development of human resources

In addition to the NATO visiting scientist, one graduate and two undergraduate students were also involved in the project. The research has so far resulted in one journal and five reviewed conference papers and one Master of Science thesis. The research also resulted in software that has been used in conducting the preliminary investigation of the current proposal. The collaboration between the engineering research team and the University of Maine Core DNA sequencing facility provided an opportunity for the team to also become familiar with the biological aspects of DNA data preparation and processing.

e. Data, samples, physical collections and other related research products

Thanks to the DNA Sequencing Core Facility at University of Maine, we get a steady source of DNA sequencing data produced by the ABI Prism 373 machine. For our preliminary studies, we randomly selected 22 groups of data; each group contains 2 ~ 3 files which are complementary or overlapped. So, for the non-well resolved part of one file, there is a corresponding part in another file, which is well resolved. By comparing the two complementary or overlapped files, we can get high confidence value sequences of the length more than 1,000 bases. Using these high quality sequences as template, we can evaluate our software and compare the result with those of others.