3-29-2004

Accurate DNA Base Caller

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

Final Report for Period: 03/2001 - 02/2004
Submitted on: 03/29/2004
Principal Investigator: Musavi, Mohamad T.
Award ID: 0090738
Organization: University of Maine
Title:
An Accurate DNA Base Caller

Project Participants

Senior Personnel

Name: Musavi, Mohamad
Worked for more than 160 Hours: Yes
Contribution to Project:
Project Director, supervising the entire project, providing technical expertise, supervising students, evaluting and testing the software.

Name: Van Beneden, Rebecca
Worked for more than 160 Hours: Yes
Contribution to Project:
Providing the biological expertise and DNA sequencing data and expertise.

Name: Ressom, Habtom
Worked for more than 160 Hours: Yes
Contribution to Project:
Providing neural network and fuzzy logic expertise, evaluating the software

Name: Domnisoru, Cristian
Worked for more than 160 Hours: Yes
Contribution to Project:
Developing algorithms for the base calling engine, supervising students, testing and evaluating the software.

Post-doc

Graduate Student

Name: Varghese, Rency
Worked for more than 160 Hours: Yes
Contribution to Project:
Developing and implementing fuzzy systems for confidence values of the base calls.

Name: French, Brian
Worked for more than 160 Hours: Yes
Contribution to Project:

Undergraduate Student

Name: Dawood, Jehad
Worked for more than 160 Hours: Yes
Contribution to Project:
Windows Programming of the base calling software (TraceTools)

Name: McNally, Ceara
Worked for more than 160 Hours: Yes
Contribution to Project:

Data preparation and processing, developing 'ground truth' data, testing TraceTools and providing error rates for TraceTools, ABI, and Phred Software

Technician, Programmer

Name: Toothaker, Mike

Worked for more than 160 Hours: Yes

Contribution to Project:

Managing Windows programming of TraceTools

Other Participant

Name: Natarajan, Padma

Worked for more than 160 Hours: Yes

Contribution to Project:

Providing quality assurance and developing help menue for TraceTools

Research Experience for Undergraduates

Organizational Partners

Other Collaborators or Contacts

We received data from Dr. Ralph Dean of Fungal Genomics Laboratory, North Carolina University.

Activities and Findings

Research and Education Activities: (See PDF version submitted by PI at the end of the report)

Findings: (See PDF version submitted by PI at the end of the report)

Training and Development:

Training and Development -First Year

During the past year the following students were supported with funds from this grant:

À One undergraduate student working part time during the entire academic year and 6 weeks during the summer of 2001. The rest of the summer she was working in our laboratory in the same project being supported by a NSF REU grant. The task was to develop the database of DNA sequences for evaluation of the software.
À Two graduate students working on the development of the methodologies and software.

Through the research conducted under this grant, The University of Maine Intelligent Systems Laboratory is able, for the first time, to offer a course in Computational Biology this coming fall. A description of the course is given below.

Course title: Algorithms in Computational Biology
Course outline:
Introduction to Molecular Biology (for computer scientists);
Restriction Maps
Cloning and Clone Libraries
Physical Genome Maps
Sequencing and Base Calling
Assembly

Sequence Analysis
Sequence Alignment Algorithms
Multiple Sequence Alignment
Probability and Statistics for Sequence Alignment
The course will be offered at a graduate level and it is intended for students in Computer Science, Computer Engineering, and Biology majors.

Training and Development -Second Year

À Two graduate students were supported and trained. They were responsible for the development of the visual interface of the software and the development and integration of the base calling techniques and confidence value.
À One undergraduate student was hired and trained. She was responsible for creating the database of sequences for evaluation of the methodology.
À One postdoc was employed full time; He was in charge of developing the algorithms and data processing techniques, provident student advice, and directly working on the software development.
À In addition to the PI, two other faculty were involved in the research for providing guidance and expertise to students, managing the project, preparing reports, and papers.
. Three teachers and 5 students from the Maine School of Science and Mathematics (MSSM) were involved in this projects.

Training and Development -Final

In addition to the PI and the three Co-PIs, this NSF grant provided opportunities for several individuals to be involved and trained in the combined research area of informatics and genomics. These are:
À Two graduate students,
À Two undergraduate students,
À One computer system developer (BS),
À One research associate (M.S.); she is now a member of the University of Maine Intelligent Systems Laboratory research team and is a Co-PI in a pending NSF proposal,
À One research faculty (Ph.D.); he is now a tenure-track professor in the bioinformatics area in a different university,
À Six high school students, and
À Three high school teachers.

Outreach Activities:

Outreach Activities -First Year

A special attention was devoted to the attraction of young students into careers in engineering and especially in the area of this project. A high school in Maine (Maine School of Science and Mathematics - MSSM) was identified as a priority target. MSSM is a public boarding school with a strong curriculum in science (mathematics, physics, computer science and biology). MSSM is a magnet for Maine high school students and its students are usually taking college level courses. It attracts the most talented high school students in Maine. Our intention was to involve MSSM students in our current research project with the hope of stimulating more students to embrace careers in bioinformatics.

On May 5th, 2002 one of the Co-PIs made a presentation of the project at MSSM to a group of 7 students and 3 teachers. The algorithmic approach to base calling and two research tasks from the project were outlined for possible approach by the students and the teachers. Currently, three students from MSSM are working part time during this summer on the project.

To complement these activities, a supplement to the current project was granted under Research Experience for Teachers (RET). This supplement will provide the funding for directly involving three teachers from MSSM during the next academic year and to attract more students.

A complementary source of funding for similar activities is also being considered.

Outreach Activities - Second Year

As a direct result of the research conducted under this project we participated in co-teaching a graduate level course at the University of Maine. The course title is 'Graduate Topics in Mathematics: Computational Methods in Genomic.' Our contribution was the preparation of 6 lectures for the course. The course topics are available at http://germain.umemat.maine.edu/faculty/crook/courses/mat500/schedule.html. Approximately 30 graduate students and researchers from the University of Maine, the Jackson Laboratory, and Maine Medical Center Research Institute attended the course. The 6 lectures were presented by the Co-PI's of this project Cristian Domnisoru and Habtom Ressom according to the following schedule:

À Cristian Domnisoru's lectures: Approaches for Enhanced Sequence Reads, Pair wise Alignment--Dynamic Programming, Multiple Sequence Alignment, Gene Finding and Protein Structure.
À Habtom Ressom's lecture: Microarray Data Analysis: Other Methods.

Under the NSF-RET supplemental award, a collaborative effort with the Maine School of Science and Mathematics (MSSM) was initiated. This started at the end of last report period and was continued during 2002-2003.

During the summer of 2002 three teachers and three students from MSSM were directly involved in this project. They became familiar with different aspects of the project and learned different steps in the preparation of the databases. They were then involved in editing sequence data files.

During the fall 2002, due to the workload required in the school, the students had to spread their effort over a longer period of time. Only two students were involved. They worked under the supervision of the three teachers involved in the project. One of the project Co-PIs made two trips to Limestone, Maine, where the school is located, to present new findings in the project and coordinate the project. The results from the research conducted during the fall were in the area of data fitting and using numerical derivatives instead of the signal variation for calculations. The students involved were able to understand the algorithms used, to modify the C++ code and to propose new methods.

Journal Publications

French, B.D., C. Domnisoru, H. Ressom and M. T. Musavi, "Confidence Value Predication of Called Genetic Bases Using a Fuzzy Predication System", Proceedings of the International Conference on Mathematic and Engineering Techniques in Medicine and Biological Sciences (METMBS), Las Vegas, NV, , CSREA Press, p. 203, vol. Vol. I, (2002). Published,

McNally, C., C. Domnisoru, M. T. Musavi, "Building a DNA Database to Compare the Accuracy of Base Calling", Proceedings of The 2002 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences, METMBS'02: Las Vegas, USA, Vol. I, CSREA Press, p. 217, vol. I, (2002). Published,

Domnisoru, C., and M.T. Musavi, "Mobility Shift Modeling for DNA Sequencing Data", Proceedings of IASTED International Symposium Modelling and Simulation, p. 336, vol. , (2001). Published,

Domnisoru, C., X. Zhan, and M.T. Musavi, "Cross-Talk Filtering in Four Dye Flourescence-based DNA Sequencing", Electrophoresis, p. 2983, vol. 21, (2000). Published,

Musavi, M.T., Domnisoru, C., Natarajan, P., Varghese, R., Toothaker, M., Dawood, J., Ressom, H., Van Beneden, R., Singer, P., "TraceTools: a new DNA basecaller", Bioinformatics, p. , vol. , ( ). To be submitted,

Varghese, R., M.T. Musavi, & R. Ressom, "A fuzzy confidence value for DNA bases", FUZZ-IEEE, p. , vol. , ( ). Accepted,

Domnisoru, D. and M.T. Musavi, "Filtering Technique for Fluorescene-Based DNA Sequencing Data", Proceedings of the 2000 IEEE Canadian Conference on Electrical and Computer Engineering, p. 498, vol. , (2000). Published,

Domnisoru, D. and M.T. Musavi, "Mechanical Shift and Base Spacing Modeling and Compensation for DNA Sequencing", Proceedings of the International Association of Science and Technology for Development (IASTED) Conference on Modeling and Simulation, p. 470, vol. , (2000). Published,

Books or Other One-time Publications

Domnisoru, D. & M.T. Musavi, "Method for reducing cross-talk within DNA data", (2003). United States Patent, Published
Bibliography: United States Patent # 6,598,013

Web/Internet Site

URL(s):
www.intsys.maine.edu/AccurateDNA
Description:
The site provides information on the project

Other Specific Products

Product Type:

Software (or netware)

Product Description:

The methodologies developed under this project have been used in a DNA base calling software, named TraceTools. TraceTool will accept any ABI 3700 DNA file and will process it for base calling. The file outcome is the base calls with corresponding confidence value for each call. The software can be run on Windows environment and in addition to base calling has some other useful functionality like showing the original raw data, printing, saving, etc.

United States Patent Office has approved and issued a patent for a section of the methodology used in TraceTool under USPTO #6,598,013

Sharing Information:

TraceTools is available from the project website at www.intsys.maine.edu/AccurateDNA.htm under downloads.

Contributions

Contributions within Discipline:

The main contribution of this project is the development of advanced mathematical methodologies and corresponding software for accurate DNA base calling. The project has developed mathematical and computer engineering knowledge to support conversion of biological data (DNA files) into meaningful and accurate information (sequences) for use by the genomic community. In doing so, the project has contributed in many ways as listed in the following categories.

Contributions within Discipline - 1st Year

A technique for DNA data preprocessing was created. This includes:
À Mechanical shift correction for files created with machines recording the four signals at different moments in time;
À Cross-talk removal using a novel approach based on the differences between successive measurements rather than the amplitude of the signals;
À Mobility shift correction;
À Baseline adjustment of the signals;
À Note that a critical aspect of the proposed technique is the order in which the above processing steps are applied. Previous work in the field indicated always the baseline adjustment step as being done before the cross-talk. We have shown our proposed order assures a much better quality of the data.

An algorithm for base calling was proposed and implemented. The algorithm:
À Creates a list of peak candidates based on the amplitude and shape of the signals (or peakness);
À Uses the peak candidates to build a base spacing model; We have found there is a strong correlation between the type of any two consecutive bases and the spacing between the corresponding peaks in the sequencing signal;
À Starts with the base spacing model and using features of the signals (amplitudes and peakness), makes the base calls.

A fuzzy-logic system for assessing base call confidence values was implemented. For each base call, two confidence values are calculated:

À The confidence that a base call should be made in the given position. This value is related to the base spacing. The further apart from the spacing model a given call is, the lower the confidence;
À The confidence that a given base call was correctly made. This value is related to the amplitudes and peakness of the peaks competing for the respective base call;
À Note that our system can also produce confident values for the second candidate, thus helping identify eventual heterozygotes.

Contributions within Discipline - 2nd Year

Additional results for DNA sequencing data preprocessing were obtained. This includes:
À Adaptive signal processing for the raw data signals;
À Implementation of a new technique for matrix detection for cross-talk filtering based on the Independent Component Analysis method;
À Mobility shift correction using chemistry dependent parameters;
À Discovery of a model for the peak amplitudes of consecutives bases. The model depends on the type of the consecutive bases in a similar manner to the base spacing model does.

We continued the implementation and tuning of the proposed base-calling algorithm. The algorithm:
À Creates a list of peak candidates based on the amplitude and shape of the signals (peakness),
À Uses the peak candidates to build a base spacing model. It was found that there is a strong correlation between the type of any two consecutive bases and the spacing between the corresponding peaks in the sequencing signal, and
À Starts with the base spacing model and using features of the signals (amplitudes and peakness) and makes the base calls.

Contributions within Discipline - Final

This project has contributed by developing:
À Several signal processing programs for smoothing and preprocessing of DNA data,
À A base space prediction strategy for accurate prediction of bases,
À Windows based DNA base calling software, named TraceTools,
À Confidence values for bases called based on fuzzy system,
À Better DNA accuracies and confidence values than ABI and Phred, and
À A database of raw and correct DNA sequences consisting of more than 3300 files.

Contributions to Other Disciplines:

Contributions to other disciplines - 1st Year

The simple cross-talk filtering technique developed has the potential to be applied to other problems where, in general, multi spectral data is recorded using sensors with overlapping spectra. One such problem is the data recording for gene microarray experiments, where two sensors are recording overlapping spectra. Although in the current technology this overlap is supposed to be minimum, from our experience with ABI sequencing data we consider there are improvements that can be made.

Similar improvements can be made to extracting secondary structure from multiple spectra using nuclear magnetic resonance (NMR) spectroscopy experiments.

Contributions to other disciplines - 2nd Year

The results we obtained with our experimentation with the Independent Component Analysis technique have the potential to be applied to other problems where, in general, multi spectral data is recorded using sensors with overlapping spectra. One such problem is the data recording for gene microarray experiments, where two sensors are recording overlapping spectra. Although in the current technology this overlap is supposed to be minimum, from our experience with ABI sequencing data we consider there are improvements that can be made.

Contributions to other disciplines - 2nd Year

Same as previous years.

Contributions to Human Resource Development:

Contributions to Human Resources Development -1st Year

À Two graduate students were supported and trained. They were responsible for the development of the visual interface of the software and the development and integration of the base calling techniques and confidence value.
À One undergraduate student was hired and trained. She was responsible for creating the database of sequences for evaluation of the methodology.
À One postdoc was employed full time; He was in charge of developing the algorithms and data processing techniques, provident student advice, and directly working on the software development.
À In addition to the PI, two other faculty were involved in the research for providing guidance and expertise to students, managing the project, preparing reports, and papers.

Contributions to Human Resources Development -2nd Year

À One graduate student was supported in part for the year 2002. He was in charge of the software development and in particular with the confidence value calculations,
À One undergraduate student was part time employed during the entire academic year 2002-2003. She was in charge of the database preparation;
À One research faculty was supported for the entire academic year 2002-2003. He worked on the development of the algorithms and supervision of the student for software implementation of the methodology.
. Three Teachers and 5 higher school students from Maine School of Science and Mathematics participated in the project.

Contributions to Human Resources Development -Final

In addition to the PI and the three Co-PIs, this NSF grant provided opportunities for several individuals to be involved and trained in the combined research area of informatics and genomics. These are:
À Two graduate students,
À Two undergraduate students,
À One computer system developer (BS),
À One research associate (M.S.); she is now a member of the University of Maine Intelligent Systems Laboratory research team and is a Co-PI in a pending NSF proposal,
À One research faculty (Ph.D.); he is now a tenure-track professor in the bioinformatics area in a different university,
À Six high school students, and
À Three high school teachers.

Contributions to Resources for Research and Education:

N/A

Contributions Beyond Science and Engineering:

N/A

Categories for which nothing is reported:

Organizational Partners

I. ACTIVITIES AND FINDINGS

A. Research and Education Activities

Two major activities were conducted. These are:

- The development of an accurate base calling technique and its associated software tool and
- The development of a database of DNA sequences for testing the base caller.

The preliminary investigation had been conducted using data from gel based sequencing machines (ABI 377). Given the gradual replacement of those machines with newer capillary electrophoresis based machines, it was decided to redirect the project effort on the development of a base caller especially calibrated for ABI 3700 sequencing machines. The major goal is to achieve an error rate of less than 1% that will provide superior results to the current 2-3% error rate provided by the ABI and Phred base calling software. Another goal is to increase the length of the read sequences obtained with a given accuracy.

The result of this study has been incorporated in a new base calling software called *TraceTool*. The software allows simultaneous view of both raw and processed data (traces). The need for viewing both raw and processed data stems from the research finding. The ABI processed data is at times altered and an experienced sequencing operator could make more confident base calls just by having access and viewing the raw data. The developed trace viewer has been completed for its basic functionality. It opens ABI raw data files, displays the raw data and the processed data in two separate windows and has the basic functionality of the Windows software. In addition, it is designed to call the proposed base calling software from within the Windows environment and to display the results. The software was developed in Visual C++. In Figure 1 a screen capture of TraceTool is presented. The screen is divided in two windows. The upper part contains the raw data, while the bottom one displays our processed data. In addition, for each base call a small rectangle indicates with green a visual representation of the confidence associated with the respective base call.

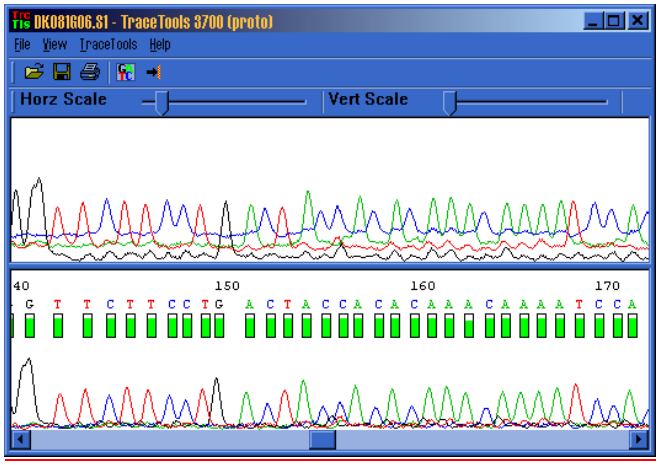


Figure 1. Screen capture of TraceTool

A special attention was devoted to the calculation of the confidence values for each base call. A fuzzy system was implemented and a graduate thesis (to be completed by August 2002) was devoted entirely to this problem.

For evaluation of the proposed techniques, there was a need for a large database of raw sequencing data along with the corresponding correct bases. In addition and for comparison, the ABI and Phred called bases of the data had to be also created and added to the database. This database has been created and is ready for evaluation of the proposed methodology and software.

B. Findings

The followings are some of the findings of this research so far.

1. The new ABI capillary electrophoresis machines have similar data streams as the gel-based machines for which the project was initially designed. Although the cross-talk effect is significantly reduced, it is still present, and a proper filtering technique is providing similar improvements as in the gel-based machines.
2. The cross-talk filtering technique based on the derivative of the signals rather than the signals themselves proved to be sound and reliable. We appreciate that this technique can be expanded to other multi-spectral interference separation problems.
3. The spacing between consecutive bases is a valid feature to help in predicting the base calls.
4. Acquiring a large database of raw DNA data and their corresponding correct sequence (ground truth) for testing the base calling software proved to be very time consuming. The difficulty stems from the fact that, the proposed technique works with raw DNA information that is available in the ABI raw trace files. While, in order to save space, sequencing centers do not keep the original ABI raw trace files and prefer to convert the data into the smaller *scf* format. Repeated inquiries into the NHGRI and other sequence databases were unsuccessful in providing any raw trace data file and its corresponding ground truth. Consequently, we were forced to develop our own database for testing. This resulted in an important time spent in acquiring data files and making the corrections to ensure correct base calls. We received the data files from the Integrated Fungal Research Center, North Carolina State University.
5. Because of the limitations of the current base callers, the ABI capillary electrophoresis machines are unable to make reasonable base calls to the full extent provided by the information content available in the streams of data. Unlike the gel-based ABI machines where data was saved well after its quality deteriorated, the new machines do not save data past the point where its proprietary software can make confident base calls. Consequently, in order to make use of our algorithm to its full extent and show that the proposed technique works even where ABI fails, there is a need to obtain an extended length of data from the manufacturer. This, however, might not be possible.

C. Training and Development

During the past year the following students were supported with funds from this grant:

- One undergraduate student working part time during the entire academic year and 6 weeks during the summer of 2001. The rest of the summer she was working in our laboratory in the same project being supported by a NSF REU grant. The task was to develop the database of DNA sequences for evaluation of the software.
- Two graduate students working on the development of the methodologies and software.

Through the research conducted under this grant, The University of Maine Intelligent Systems Laboratory is able, for the first time, to offer a course in Computational Biology this coming fall. A description of the course is given below.

Course title: Algorithms in Computational Biology
Course outline:

Introduction to Molecular Biology (for computer scientists);
Restriction Maps
Cloning and Clone Libraries
Physical Genome Maps
Sequencing and Base Calling
Assembly
Sequence Analysis
Sequence Alignment Algorithms
Multiple Sequence Alignment
Probability and Statistics for Sequence Alignment

The course will be offered at a graduate level and it is intended for students in Computer Science, Computer Engineering, and Biology majors.

D. Outreach Activities

A special attention was devoted to the attraction of young students into careers in engineering and especially in the area of this project. A high school in Maine (Maine School of Science and Mathematics - MSSM) was identified as a priority target. MSSM is a public boarding school with a strong curriculum in science (mathematics, physics, computer science and biology). MSSM is a magnet for Maine high school students and its students are usually taking college level courses. It attracts the most talented high school students in Maine. Our intention was to involve MSSM students in our current research project with the hope of stimulating more students to embrace careers in bioinformatics.

On May 5th, 2002 one of the Co-PIs made a presentation of the project at MSSM to a group of 7 students and 3 teachers. The algorithmic approach to base calling and two research tasks from the project were outlined for possible approach by the students and

the teachers. Currently, three students from MSSM are working part time during this summer on the project.

To complement these activities, a supplement to the current project was granted under Research Experience for Teachers (RET). This supplement will provide the funding for directly involving three teachers from MSSM during the next academic year and to attract more students.

A complementary source of funding for similar activities is also being considered.

II. PUBLICATIONS AND PRODUCTS

A. Publications

Domnisoru, C. and M. T. Musavi, "Mobility shift Modeling for DNA Sequencing Data", Proceedings of the MS'2001 Modeling and Simulation, IASTED International Conference, Pittsburgh, Pennsylvania, USA, pp.336-342, May 16-18, 2001.

French, B.D., C. Domnisoru, H. Ressom and M T. Musavi, "Confidence Value Predication of Called Genetic Bases Using a Fuzzy Predication System," The 2002 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences, METMBS'02: Las Vegas, USA, June 24-27, 2002.

McNally, C., C. Domnisoru, M. T. Musavi, "Building a DNA Database to Compare the Accuracy of Base Calling," The 2002 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences, METMBS'02: Las Vegas, USA, June 24-27, 2002.

Domnisoru, C., "Modeling and Removal of "Reflective" Phenomena in Electrophoresis Sequencing Data," IASTED International Conference on Modeling and Simulation, Marina del Rey, California, May 13-15, 2002.

B. Books and Other One-Time Publications

N/A

C. Other Specific Products

A database of DNA raw files and their corresponding correct sequence (ground truth) was created. The database is needed for comparing the accuracy of the proposed base-calling program with other base calling software (ABI and Phred). The objective was to create a database of correct DNA sequences accompanying the corresponding ABI raw trace data. Constructing this database required a database of raw ABI data files, a large contig containing those sequences assembled within, and including the property of having

the correct bases for sequences. The database that has 1,139 files obtained from the North Carolina State University, Integrated Fungal Research Center. The files contain roughly 7 hundred thousand bases. Extracting the ABI called sequences from each raw data file in the database revealed they contained some errors. In order to correct for those errors, the associated contig was used and each of the individual sequences identified within the contig. The sequence was corrected based on the assumption that the contig, being obtained through a 10 fold overlapping sequencing project, will provide a higher confidence per base. In addition to the ABI raw data files and the correct sequences, the ABI called bases and Phred called bases were also added to the database for comparison with our base calling technique and any other that might be developed in the future.

A software called "*TraceTool*" was created by two graduate students for integration of the results of this project. The software can visualize the sequencing data and call bases. The visualization option is to show both the raw trace data (signal recorded by the sequencing machines) and the processed trace data. There is another similar software, Chromas by Technelysium Pty Ltd, Australia, that can also visualize the data but it can only show the processed data. Recognizing the need to offer the sequence operator the opportunity to see the original raw data recorded by the sequencing machine, together with the processed data, this tool was created. It is important to be capable of viewing both types of data simultaneously for better identification of bases, if in doubt. In addition, the software also integrates the proposed methodologies for calling of the bases along with the confidence value for each base.

The University of Maine has filled a U.S. Patent application in August 2001 on the techniques used in TraceTool. The patent application covers the data processing techniques, including the base line adjustment, cross-talk filtering, and noise filtering.

III. CONTRIBUTIONS

A. Contributions within Discipline

A technique for DNA data preprocessing was created. This includes:

- Mechanical shift correction for files created with machines recording the four signals at different moments in time;
- Cross-talk removal using a novel approach based on the differences between successive measurements rather than the amplitude of the signals;
- Mobility shift correction;
- Baseline adjustment of the signals;
- Note that a critical aspect of the proposed technique is the order in which the above processing steps are applied. Previous work in the field indicated always the baseline adjustment step as being done before the cross-talk. We have shown our proposed order assures a much better quality of the data.

An algorithm for base calling was proposed and implemented. The algorithm:

- Creates a list of peak candidates based on the amplitude and shape of the signals (or peakness);
- Uses the peak candidates to build a base spacing model; We have found there is a strong correlation between the type of any two consecutive bases and the spacing between the corresponding peaks in the sequencing signal;
- Starts with the base spacing model and using features of the signals (amplitudes and peakness), makes the base calls.

A fuzzy-logic system for assessing base call confidence values was implemented. For each base call, two confidence values are calculated:

- The confidence that a base call should be made in the given position. This value is related to the base spacing. The further apart from the spacing model a given call is, the lower the confidence;
- The confidence that a given base call was correctly made. This value is related to the amplitudes and peakness of the peaks competing for the respective base call;
- Note that our system can also produce confident values for the second candidate, thus helping identify eventual heterozygotes.

B. Contributions to other disciplines

The simple cross-talk filtering technique developed has the potential to be applied to other problems where, in general, multi spectral data is recorded using sensors with overlapping spectra. One such problem is the data recording for gene microarray experiments, where two sensors are recording overlapping spectra. Although in the current technology this overlap is supposed to be minimum, from our experience with ABI sequencing data we consider there are improvements that can be made.

Similar improvements can be made to extracting secondary structure from multiple spectra using nuclear magnetic resonance (NMR) spectroscopy experiments.

C. Contributions to Human Resources Development

- Two graduate students were supported and trained. They were responsible for the development of the visual interface of the software and the development and integration of the base calling techniques and confidence value.
- One undergraduate student was hired and trained. She was responsible for creating the database of sequences for evaluation of the methodology.
- One postdoc was employed full time; He was in charge of developing the algorithms and data processing techniques, provident student advice, and directly working on the software development.
- In addition to the PI, two other faculty were involved in the research for providing guidance and expertise to students, managing the project, preparing reports, and papers.

D. Contributions to Resources for Research and Education

N/A

E. Contributions Beyond Science and Engineering

N/A

Findings – 1st Year

The followings are some of the findings of this research so far.

1. The new ABI capillary electrophoresis machines have similar data streams as the gel-based machines for which the project was initially designed. Although the cross-talk effect is significantly reduced, it is still present, and a proper filtering technique is providing similar improvements as in the gel-based machines.
2. The cross-talk filtering technique based on the derivative of the signals rather than the signals themselves proved to be sound and reliable. We appreciate that this technique can be expanded to other multi-spectral interference separation problems.
3. The spacing between consecutive bases is a valid feature to help in predicting the base calls.
4. Acquiring a large database of raw DNA data and their corresponding correct sequence (ground truth) for testing the base calling software proved to be very time consuming. The difficulty stems from the fact that, the proposed technique works with raw DNA information that is available in the ABI raw trace files. While, in order to save space, sequencing centers do not keep the original ABI raw trace files and prefer to convert the data into the smaller *scf* format. Repeated inquiries into the NHGRI and other sequence databases were unsuccessful in providing any raw trace data file and its corresponding ground truth. Consequently, we were forced to develop our own database for testing. This resulted in an important time spent in acquiring data files and making the corrections to ensure correct base calls. We received the data files from the Integrated Fungal Research Center, North Carolina State University.
5. Because of the limitations of the current base callers, the ABI capillary electrophoresis machines are unable to make reasonable base calls to the full extent provided by the information content available in the streams of data. Unlike the gel-based ABI machines where data was saved well after its quality deteriorated, the new machines do not save data past the point where its proprietary software can make confident base calls. Consequently, in order to make use of our algorithm to its full extent and show that the proposed technique works even where ABI fails, there is a need to obtain an extended length of data from the manufacturer. This, however, might not be possible.

Findings – 2nd Year

The followings are the findings of this research for the second year.

We started to process a large database (4300 files) received from the Fungal Genomics Laboratory, North Carolina University. From processing of these files several general problems appeared, problems that we had not taken into account previously. These are:

1. The signal level in different files varies significantly. While for some files the signal to noise ratio is very low, hence, making the noise filtering unnecessary, for others the ratio is very high. To address this problem we created an adaptive filtering system. In this way, the files with a high signal to noise ratio were treated differently.
2. The cross-talk filtering system we proposed proved to work correctly with one exception. This exception is when the cross-talk matrix has two or more components in one line (or column) of similar (and high) magnitude. To address this problem we reconsidered the cross-talk filtering system and find that our initial approach was not correct in the general case. We also identified the similarity between our problem and the "cocktail party's problem" and therefore adapted our algorithm for cross-talk filtering to use advanced results recently published in this area. We implemented the Independent Component Analysis technique (Hyvärinen, A., Oja, E., "Independent Component Analysis: Algorithms and Applications," *Neural Networks*, 13(4-5): 411-430, 2000) for detecting the cross-talk matrix. However, the use of the derivative of the signals as we proposed earlier still holds as a major contribution to this filtering problem.
3. On the mobility shift correction, the four types of fragments do not migrate with the same speed. As long as the peaks do not change their natural order when they pass through the detector area, the uneven spacing will be captured by the base spacing model and even improve the base calling. However, in the case of CE data, at the beginning of the sequence there is a short region where the peaks are changing their order due to a different mobility. We previously proposed a technique to automatically detect and compensate for this effect. However, it appears simpler to apply a classic mobility shift correction correlated with the sequencer and the chemistry used. There are a few parameters that need to be estimated for each chemistry used. Note that we are not looking for a precise mobility correction, but rather an approximate one that requires the peak to maintain the natural order. That's why, the estimation of the parameters for the mobility shift correction should be a simple task and it will be performed as part of the maintenance and update of the software.
4. On the peak amplitude prediction, we found out that there is a strong correlation between the types of consecutive bases and the amplitude ratio of the corresponding peaks. In other words, we found an additional feature for predicting the next base call besides the base spacing. For instance, in our processed files more than 90% of the cases, for a C-G-T sequence of bases, the ratio between the amplitude of the G peak and the T peak was less than 1 (with an average of 0.5). Also, for a sequence C-G-A the ratio between the amplitude of

the G peak and the A peak was greater than 1 more than 90% of the time, with an average of 1.5. This shows that if we already know the bases starting from the good region, (in the example above C-G), in order to predict the next base, we can not only check the base spacing model, but also we can build a peak amplitude model and use this information for better predicting the base call. Note that this is a very important result that we have not implement in the software yet.

Findings – Final

Specific findings of this research have been listed in the finding of the 1st and 2nd years. After the completion of the testing and evaluations in the final year, it was also found out that:

1. The performance of the current popular DNA sequencing software can indeed be improved as indicated in the results of this research. See the *Activities* section.
2. A reliable confidence value for bases called, such as the one developed in this project, can be effectively used to identify the areas of high error and consequently be used for further base calling improvement.
3. Availability of ground truth (correct DNA) data is critical for evaluation purposes. It is also noted that creation of such datasets is not a trivial task due to the fact that: 1) not many genomic centers preserve the raw data and 2) even the correct data may contain errors.
4. The proprietary nature of the DNA raw data, which is mostly created by the ABI sequencing machines, has made it more difficult for independent sequencing software to use the raw data easily.

Activities – 1st Year

The goal of this project is to develop a new base calling technique to improve the efficiency of the DNA sequencing process. Towards achieving this goal, two major activities were conducted:

- The development of an accurate base calling technique and its associated software tool and
- The development of a database of DNA sequences for testing the base caller.

The preliminary investigation had been conducted using data from gel based sequencing machines (ABI 377). Given the gradual replacement of those machines with newer capillary electrophoresis based machines, it was decided to redirect the project effort on the development of a base caller especially calibrated for ABI 3700 sequencing machines. The major goal is to achieve an error rate of less than 1% that will provide superior results to the current 2-3% error rate provided by the ABI and Phred base calling software. Another goal is to increase the length of the read sequences obtained with a given accuracy.

The result of this study has been incorporated in a new base calling software called *TraceTool*. The software allows simultaneous view of both raw and processed data (traces). The need for viewing both raw and processed data stems from the research finding. The ABI processed data is at times altered and an experienced sequencing operator could make more confident base calls just by having access and viewing the raw data. The developed trace viewer has been completed for its basic functionality. It opens ABI raw data files, displays the raw data and the processed data in two separate windows and has the basic functionality of the Windows software. In addition, it is designed to call the proposed base calling software from within the Windows environment and to display the results. The software was developed in Visual C++. In Figure 1 a screen capture of TraceTool is presented. The screen is divided in two windows. The upper part contains the raw data, while the bottom one displays our processed data. In addition, for each base call a small rectangle indicates with green a visual representation of the confidence associated with the respective base call.

Figure 1. Screen capture of TraceTool

A special attention was devoted to the calculation of the confidence values for each base call. A fuzzy system was implemented and a graduate thesis (to be completed by August 2002) was devoted entirely to this problem.

For evaluation of the proposed techniques, there was a need for a large database of raw sequencing data along with the corresponding correct bases. In addition and for comparison, the ABI and Phred called bases of the data had to be also created and added to the database. This database has been created and is ready for evaluation of the proposed methodology and software.

Activities – 2nd Year

The project's two major activities in the development of the methodologies and the creation of the database, as reported in the first report were continued.

Research for improving the algorithm for DNA base calling was conducted. Several shortcomings in the original algorithm were identified and research was conducted to address the problems. Specifically:

- An adaptive noise filtering technique was implemented,
- A better method for detecting the matrix for cross-talk filtering was implemented using Independent Component Analysis,
- The mobility shift correction was found to be better taken into consideration by parameters adapted for each type of chemistry used rather than performing it automatically, and
- A correlation between the amplitudes of consecutives peaks and their respective types was discovered – adding this feature to our base spacing model will help increase the base call accuracy.

In regard to the database creation, the database of DNA sequences that were initially collected was updated to 4300 ABI data files. The original database and its updated version were acquired from the Fungal Genomic Laboratory of North Carolina State University (Professor Ralph Dean). This database not only keeps the raw data of the ABI files, which is needed in our techniques, but it also has a long contig containing the "correct" or "ground truth" base calls for all the files. Four different datasets were created from the 4300 data files and the correct contig:

- ABI dataset - contains the sequences of ABI base calls. The ABI base calls were already available in the 4300 data files and it was a matter of extracting them from the files.
- Phred dataset - contains the sequences of Phred base calls. To get these calls, Phred software was run on each file.
- TraceTools dataset - contains the sequences of base calls generated by this project.
- "Correct" or "Ground Truth" dataset – contains the correct sequences corresponding to the 4300 data files. To create the "correct" sequence for each file, the long contig of correct sequences was searched using BLAST for each file using Phred's generated base calls.

The files of the first three datasets listed above were compared against the correct dataset to measure the accuracy of base calls for ABI, Phred, and TraceTools. The reason for selecting Phred's outcome in BLAST for creation of the correct dataset was because overall Phred's has provided better performance than ABI. TraceTools couldn't have been used because it was the software under investigation. As such, it should be noted that there is a possibility that the generated correct dataset could be biased towards Phred's results.

Outreach Activities

As a direct result of the research conducted under this project we participated in co-teaching a graduate level course at the University of Maine. The course title is "Graduate Topics in Mathematics: Computational Methods in Genomic." Our contribution was the preparation of 6 lectures for the course. The course topics are available at http://germain.umemat.maine.edu/faculty/crook/courses/mat500/schedule.html. Approximately 30 graduate students and researchers from the University of Maine, the Jackson Laboratory, and Maine Medical Center Research Institute attended the course. The 6 lectures were presented by the Co-PI's of this project Cristian Domnisoru and Habtom Ressom according to the following schedule:

- Cristian Domnisoru's lectures: Approaches for Enhanced Sequence Reads, Pair wise Alignment--Dynamic Programming, Multiple Sequence Alignment, Gene Finding and Protein Structure.
- Habtom Ressom's lecture: Microarray Data Analysis: Other Methods.

Under the NSF-RET supplemental award, a collaborative effort with the Maine School of Science and Mathematics (MSSM) was initiated. This started at the end of last report period and was continued during 2002-2003.

During the summer of 2002 three teachers and three students from MSSM were directly involved in this project. They became familiar with different aspects of the project and learned different steps in the preparation of the databases. They were then involved in editing sequence data files.

During the fall 2002, due to the workload required in the school, the students had to spread their effort over a longer period of time. Only two students were involved. They worked under the supervision of the three teachers involved in the project. One of the project Co-PIs made two trips to Limestone, Maine, where the school is located, to present new findings in the project and coordinate the project. The results from the research conducted during the fall were in the area of data fitting and using numerical derivatives instead of the signal variation for calculations. The students involved were able to understand the algorithms used, to modify the C++ code and to propose new methods.

Activities – Final

The final activities of this project resulted in the development of a novel base calling software, named ***TraceTools***, with a unique user-friendly confidence value feature.

TraceTools, as shown in Figure 2 below, is Windows based software that can display both the raw (upper window) and the processed data (lower window) after making base calls. The display of raw data allows the user to view the data as recorded by the sequencing machine. When the base calls made are uncertain, this display feature would help the user make confident decisions after investigating the raw data. The base calling is done through a processing of the raw data rather than using the pre-processed results obtained by the *ABI* software. *TraceTools* also displays a confidence value associated with each base call through a color-coded rectangular bar. After calling bases, *TraceTools* can write the sequences to files in FASTA format. Other features of *TraceTools* are: exporting raw and processed data into text files, printing data and display screen, searching for particular sequences by exact match, copying and pasting, and a help menu. More features are currently being added to the system. A copy of *TraceTools* can be downloaded from the project web site at:

http://www.intsys.maine.edu/AccurateDNA.htm#Downloads

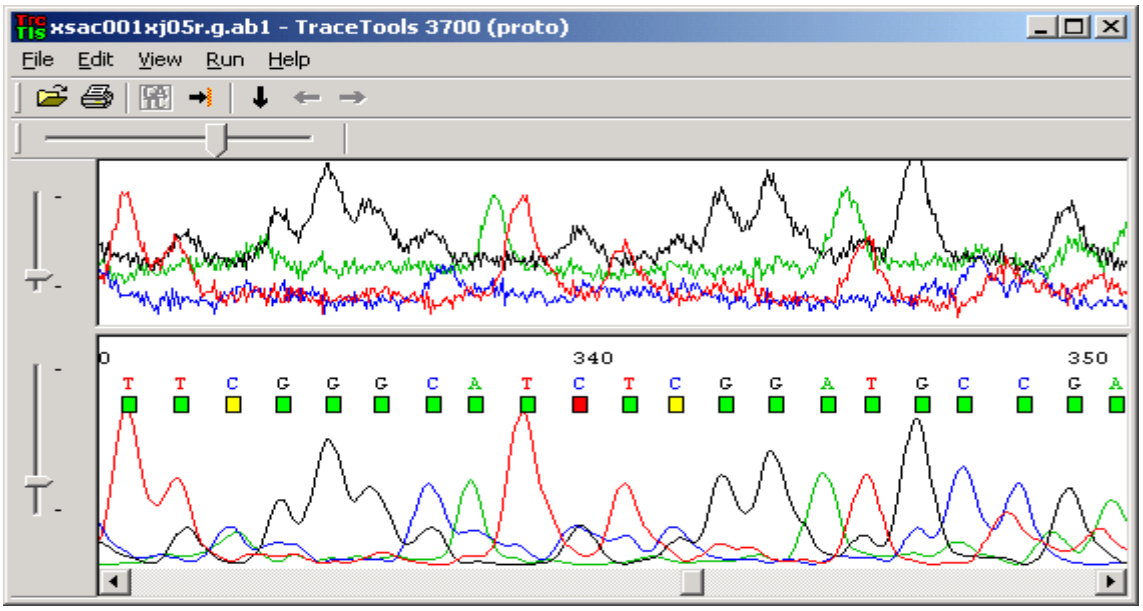


Figure 2: *TraceTools window showing a sequence with bases called and their confidence values indicated by color-coded bars; upper window shows unprocessed data.*

TraceTools' confidence value is a continuous value between 0 and 1, with 1 indicating the highest confidence. The values are generated by a fuzzy system. For ease of recognitions of poor calls, the confidence values are however, indicated through rectangular bars in the display. Green indicates high confidence (50% or higher). The green box is further split into three parts to indicate confidence between 50 and 100%. If just the bottom 1/3 of the bar is colored green, the

confidence value is 50% - 60%. If the bottom 2/3 is colored green, the confidence measure is 60% - 80%. A fully colored green bar indicates 80% - 100% confidence. Yellow colored bar indicates 25% to 50% confidence. Red indicates low confidence in the results obtained (25% or below). This makes it very easy for possible operator's intervention. As an example, a section of a DNA sequence is shown in the Figure 2 that contains bases with very high confidence (full green), medium confidence (yellow) and poor confidence (red).

For testing *TraceTools*, a comprehensive database of more than 3362 raw and corresponding *correct* DNA sequences, which had been prepared in the previous activities, was used. Note that the data were from ABI 3700 sequencer. The accuracy results were compared with popular base calling programs such as *Phred* and *ABI*. Several programs were developed to use *TraceTools* in a batch mode, record the results, and compare the results with the correct sequence to get the accuracy. The same was done with ABI and Phred software. The accuracy of *TraceTool* was then compared with those of ABI and Phred. Several restrictions such as the trimmed ends of sequences and the error calculation for only the best matching part prevented the application of *Blast* for comparing the outcome of base-calling with the corresponding correct sequence. Therefore, an alignment program was developed that when provided with two sequences produced the alignment of those sequences and reported accuracy as well as a visual base-by-base comparison of the files. This visual comparison allows the user to view where the files do not match and investigate the reason for mismatch.

The average of accuracy for all the files were 95.99% for *ABI*, 97.11% for *Phred*, and 97.47% for *TraceTools*. Figure 3 shows the histogram of accuracy for *ABI*, *Phred* and *TraceTools* for the accuracy range 80 –100%. Figure 4 shows the difference histogram between *TraceTools* and *ABI* and between *TraceTools* and *Phred* for the same accuracy range. As can be seen from the figures, *TraceTools* performs better than *ABI* and *Phred* for a majority of the files. Note that the overall accuracy provides an average measure of accuracy over all ranges of accuracies while Figures 3 and 4 provides histogram information in the given range. For example, Figure 3 indicates that the number of files with accuracy greater than 94% using *TraceTools* is more than *Phred* and *ABI*, while Figure 4 provides the difference in the number of files. As shown in Figure 4, *TraceTools* had about 300 more files in that range as compared to *ABI* and 100 more files as compared to *Phred*.

Our experiments with *TraceTools* confidence values also provided more accurate representation of the base call correctness as compared to *Phred* and *ABI* quality values. As an example, Figure 5 gives a screenshot of *TraceTools* where a section of a sequence is shown and the confidence values are also indicated by colored bars above the bases. Note that the numerical values for these confidence measures are available, as given in the 2nd row of Table 1. However, color-coded bars are used in the display for easy visual inspection. The 1st row of Table 1 gives the corresponding *Phred's* quality values. By looking at Figure 5, it is obvious that the measure of correctness of any base caller for calling the first T base should have the best confidence value among all the other bases. *TraceTools* has assigned a confidence value of 0.93 (out of 1), which is the highest among all other values. While *Phred*'s quality value for the same base is 12 (out of 50), which is surprisingly the lowest. Note that in *Phred,* the higher the number is, the better the base call should be. Similar observations can be made for other bases. For example, the trace data in Figure 5 clearly shows that the 2nd T base should have a better confidence value than any

of the two C bases around it. While *TraceTools* clearly shows this distinction in its confidence value, *Phred's* quality value provides exactly the opposite. This shows an inconsistency in the assignment of confidence values or quality values by *Phred.*

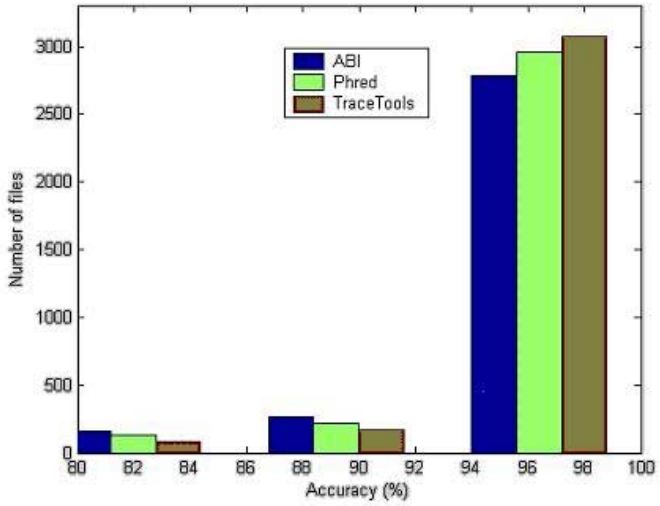


Figure 3: Histogram of accuracy for *ABI*, *Phred* and *TraceTools* for range 80-100%.

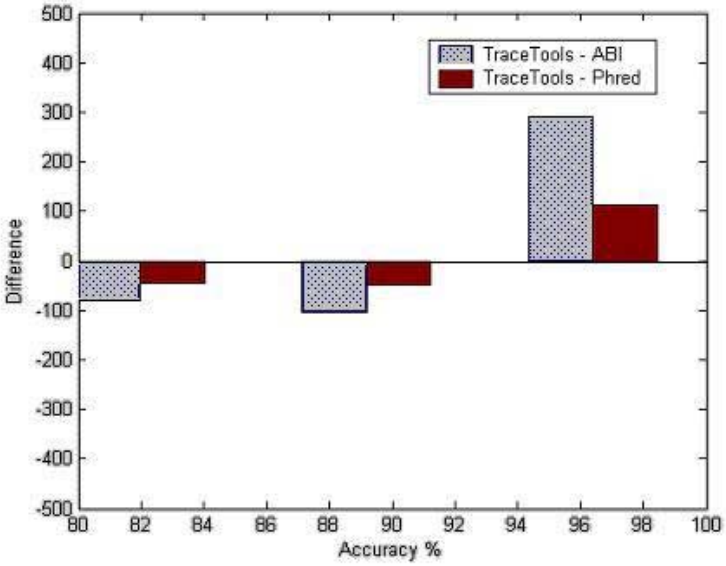


Figure 4: Difference histograms between *TraceTools* and *ABI* & *TraceTools* and *Phred*.

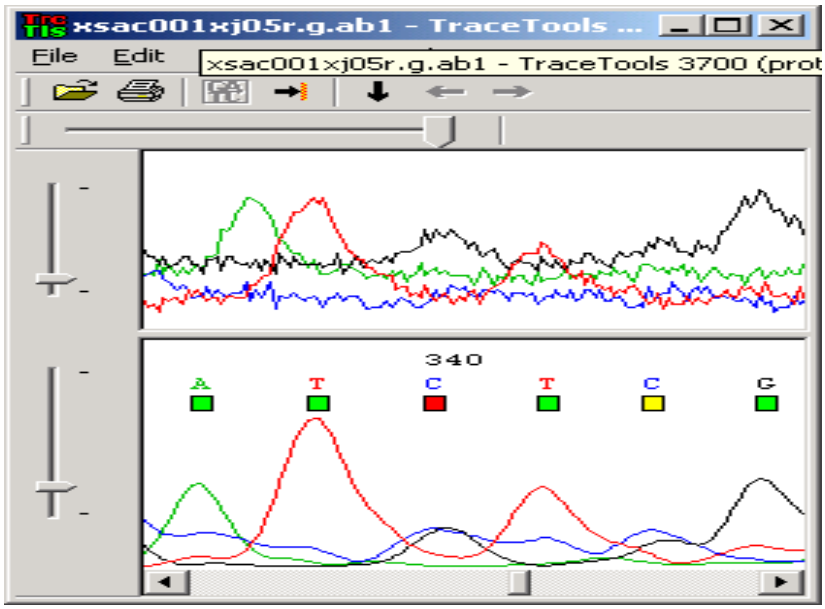


Figure 5: A screenshot of *TraceTools* for confidence value analysis.

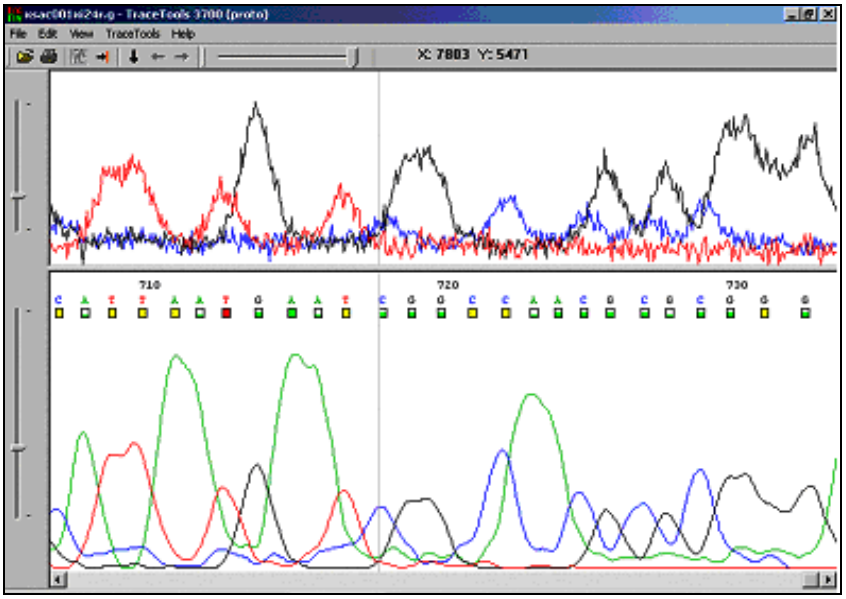
Table 1: Confidence values for *Phred* and *TraceTools* for the sequence shown in Figure 5.

	Confidence values for the bases called					
	A	**T**	**C**	**T**	**C**	**G**
Phred (Max value50)	20	12	14	13	15	22
TraceTools (Max value 1)	0.84	0.92	0.12	0.78	0.22	0.90

There are a couple of points that one needs to consider regarding the accuracy results presented above. One point is the fact that the correct contig that was used for finding the accuracy is not 100% full proof. In other words, the contig has shown to have some errors in it. The other point is that the accuracy results may be biased in the favor of *Phred* because *Phred's* generated sequences were used with *Blast* to extract the correct sequences from the contig. The errors in contig was observed in the accuracy analysis. In finding the accuracy, the traces were also analyzed and confirmed manually for a large number of files. In doing this, it was observed that in some cases, *TraceTools* made the right call, but the contig (ground truth) was incorrect. To give some examples, Figures 6 shows a screenshot of the *TraceTool*s result for the chromatogram file tested. Shown to the right of the screenshot is the part of the sequence called by *TraceTools* (bottom) and indicated by contig (top). The same is also given for ABI and Phred. As seen, *TraceTools* calls the 2nd base to be a *C* and contig indicates it as a *T*. However, analyzing the results by comparing it with the raw data shows that *TraceTools* indeed made the

correct base call. This fact is also verified by ABI and Phred, as indicated on the figure. Even though *TraceTools* makes a correct base call in cases such as the above, it is not accounted for in the calculation of its accuracy as the accuracy percentages are calculated with respect to the ground truth (contig).

Figure 7 shows an example when *TraceTools* was correct and the called base matched that of the contig while others failed. The figure shows the result of *Phred*, *ABI* and *TraceTools* for a test file (bottom row). The bases indicated by the contig are shown in the top row. The vertical lines show that a match exists between the base caller results and the contig. As evident from the results *TraceTools* makes the base call *C* similar to the contig, but *ABI* calls it a *G* and *Phred* calls it a *T*.



Contig (top row)
1-TraceTools,
2- Phred and
3- ABI
(bottom row)

```
TTGG
|  ||
TCGG
```

```
TTGG
|  ||
TCGG
```

```
TTGG
|  ||
TCGG
```

Figure 6: Screenshot of *TraceTools* results for a chromatogram file (upper window: raw data; lower window: base calls made by *TraceTools*) (Right) *TraceTools, Phred and ABI* base calls w.r.t. Contig; vertical lines show match between them.

```
Contig →      TTTCCT
              ||| ||
Phred →       TTTTCT
```

```
Contig →      TTTCCT
              ||| ||
ABI →         TTTGCT
```

```
Contig →      TTTCCT
              ||||||
TraceTools →  TTTCCT
```

Figure 7: *Phred*, *ABI* and *TraceTools* base calls w.r.t the contig for a chromatogram file.