

***TraceTools*: a new DNA basecaller**

Mohamad T. Musavi^{*1}, Cristian Domnisoru², Padma Natarajan¹, Rency Varghese¹, Mike Toothaker¹, Jehad Dawood¹, Habtom Resson¹, Rebecca Van Beneden³, Patty Singer³

*corresponding author: musavi@eece.maine.edu

Phone: (207) 581-2243; Fax: (207) 581-4531

¹Intelligent Systems Laboratory

Department of Electrical and Computer Engineering

University of Maine, 5708 Barrows Hall, Orono, ME 04473, USA

²University of St. Thomas, St. Paul, MN 55105, USA

³DNA Sequencing Facility, School of Marine Sciences

University of Maine, 351 Hitchner Hall, Orono, ME 04473, USA

Abstract

This paper presents a novel basecalling software, *TraceTools*, with a unique user-friendly confidence value feature. Because of the differences between *ABI* data pre-processing and those of this software, *TraceTools* uses the raw data generated by *ABI* machines as opposed to the *ABI* processed data. For developing and testing *TraceTools*, a comprehensive database of *correct* DNA sequences corresponding to the *ABI* raw data was constructed. The *ABI* raw trace data was obtained from North Carolina State University. This comprehensive database was used for comparing the accuracy of *TraceTools* with other popular basecalling programs such as *Phred* and *ABI*. The results of this comparison on over 3000 data files with around 750 bases per file show that *TraceTools* performs better than *ABI* and *Phred*. The confidence values also provide a reliable measure of success in the bases called.

Keywords – *ABI*, basecalling accuracy, confidence values, DNA basecalling, *Phred*, *TraceTools*.

1. INTRODUCTION

While the automated DNA basecalling is an established procedure for years, the need for increased accuracy, longer reads and increased confidence remains very important for any sequencing center. Current sequencing technology produces error rates in the order of 1.5% for a 550 base pairs read (*ABI 3700* analyzer specifications). For longer reads, the error rate is higher, reaching 2.5% or more. Many sequencing projects are underway at a lower coverage rate due to the cost associated with sequencing. This implies acceptance of a large number of gaps and a significant total gap length. In particular, these limitations are severe for Bacterial Artificial Chromosomes (BAC) sequencing implying an increased cost for sequencing and less data output for future investigations.

The technology for sequencing DNA has rapidly evolved from gel based to capillary electrophoresis (CE). Though the machines have been replaced there is hardly any change in the appearance of the data to be analyzed from a user's perspective. What a user sees is a succession of peaks of four different colors corresponding to the four bases: G, T, A, and C (Guanine-black, Thymine-red, Adenine-green, Cytosine-blue). If the peaks were clearly separated and big enough when compared to the noise at the baseline, making basecalls would be easier either through visual inspection or using automated software.

The most used sequencing systems are the *ABI* sequencing machines. In general, DNA fragments are tagged with fluorescent dyes at lengths corresponding to the number of bases in the fragment. The strands are then separated by length using electrophoresis. Individual samples to be scanned are passed through separate capillaries. A laser beam scans the strands and the reflected intensities from each of the four bases are recorded. Though the output of this physical process is affected by noise, interference between the four filters and other phenomena is less understood.

The *ABI* sequencing software for gel-based machines achieved an error rate of about 6.1% (Koonin, 1997). The *ABI PRISM 3700* produces error rates of the order of 1.5% for a 550 base pairs

read (*ABI* 3700 analyzer specifications). This is very high considering the requirements of sequencing projects. Olson (Olson and Green, 1998), Richterich (Richterich, 1998), and Ewing (Ewing et. al, 1998), have indicated a commonly accepted base-pair accuracy goal of 99.99%. Sequences with an accuracy lower than 99.90% is considered of little help in identifying single nucleotide polymorphisms in humans; see Felsenfeld et. al., 1999.

The primary alternative to *ABI*'s software is *Phred*, developed over the years by a group of contributing researchers and supported by researchers at the Department of Molecular Biology, University of Washington. *Phred* uses an advanced heuristic algorithm based on Fourier analysis and dynamic programming for base selection. *Phred* constantly outperforms the *ABI* basecaller, even for the CE machines. An excellent additional feature of *Phred* is that for each basecall it provides a quality value. This established a *de facto* standard for sequencing centers. Although *ABI* didn't provide quality values in their earlier software, they have recently integrated a quality value similar to *Phred*'s in their *ABI* 3730 machines.

The data processing steps from the raw data to called bases can be divided in two parts: *pre-processing* and *basecalling*. Pre-processing step includes adjustment for mobility shift, removal of the interference between the channels (cross-talk removal), baseline adjustment, noise filtering, and finally preparation of a normalized set of four streams of peaks for identification of the bases. As this part changes with the technology, the maker of the sequencing machine is responsible to create the software for pre-processing. In the case of *ABI*, the change from gel based to CE has clearly changed the profile of the physical system output. The computational effort was directed to create pre-processing software that will have an output similar to the one provided by the old machines. This would provide the flexibility to use existing basecalling programs, with minor modifications if necessary, for the second part of data processing.

It is clear that the overall performance of DNA base identification would depend on the accuracy of each of the two data processing steps. No matter how well-designed a basecaller would be, a less accurate pre-processing will affect the overall performance. In our previous work (Domnisoru et. al, 2000; Domnisoru and Musavi, 2003), several techniques were provided as improvements to the current *ABI* pre-processing techniques. For example, it was shown that the cross-talk removal should be performed before the base-line adjustment in order to preserve the integrity of the original data. Furthermore, the cross-talk removal is more effective if performed on the difference signal instead of the signal itself.

This paper describes and provides results on the software (*TraceTools*) that was built based on our pre-processing and basecalling algorithms. The paper is organized as follows. Section 2 describes materials and methods used for developing *TraceTools*. Section 2.1 describes the database preparation needed for obtaining the correct sequence of bases for each DNA file. Section 2.2 describes the data processing and the basecalling algorithm used by *TraceTools*. Section 2.3 provides information on assigning confidence values. Section 2.4 gives a brief description of the main features of *TraceTools*. Section 3 provides a discussion on the results obtained by testing, validation, and comparison of *TraceTools* with *Phred* and *ABI* basecallers. Section 4 provides conclusion.

2. MATERIALS AND METHODS

2.1 DATABASE PREPARATION

The database for this study requires two sets of files: raw data files and the corresponding correct sequence files. The raw data file is used as the input to the basecalling system and the correct sequence is used as the ground truth for accuracy measurement. About 4000 raw data files from *ABI* 3700 machine were obtained from North Carolina State University Fungal Genomics Lab. This center was one of only a few places that kept the raw data when saving the *ABI* files. Most centers discard the raw data and only keep the sequence information to save memory. Availability of raw data is a

requirement because *TraceTools* pre-processing steps are performed independent of *ABI* pre-processing technique. Note that *Phred* uses the *ABI* pre-processed data.

The corresponding correct sequences (ground truth) for the raw files were not readily available. Instead, a large contig of correct sequences, comprised of thousands of overlapping sequences, was available. The ground truth sequence for each raw data file had to be extracted from this large contig (McNally et. al, 2002). This contig contained the sequences of *ABI* raw files assembled within it and the information for making the correct basecalls for each sequence. The data quality of the obtained *ABI* files was predominantly good in the center portion, but poor at the beginning and end of the file. The purpose of this contig was to provide the correct basecalls for each sequence so that we could obtain a file containing the most accurate basecalls from start to end (with the exception of the primer). This correct file would enable making reliable evaluation of the basecalling program. Each base in the contig was formed from as many as five or six sequence files which had overlapped at that base position thereby providing concurrent information about the location of that base. This information from many overlapping files for a base increased the confidence of claiming that base as the correct base for that position in the sequence.

With the availability of the large contig, the next step was to search the contig for the correct sequence corresponding to each raw file in the database. There were three steps involved in finding the correct sequences. First, *Phred* basecalling software was run on each raw data file in the database and a sequence was obtained. *Phred* was used because it has demonstrated very good accuracy when tested over wide range of sequencing methods; see Richterich (Richterich, 1998) and Ewing et al. (Ewing et. al., 1998). Furthermore, *Phred* has been shown to be more accurate than *ABI* basecalling software in terms of insertion, deletion, substitution and total error. Subsequently *blast2seq* (Tatusova and Madden, 1999) was run on each *Phred* generated sequence with the large contig providing information on the alignment location within the contig. If a match was found, the corresponding sequence from the contig was saved as the correct sequence for the file (except for the primer region).

The search between the contig and the sequence file was also performed for a possible reverse complement match. A program written in C++ automated the above search process. This program initially used loops to sequentially open each file in the database. The result of this search method produced 3362 correct sequence files within the contig corresponding to 3362 raw *ABI* files. These files were used to compare the accuracy of *TraceTools* basecalls with other currently popular basecalling programs such as *ABI* and *Phred*.

Though the files found within the contig contained the correct bases for the raw data, they did not contain the primer. For an *ABI* file, a primer is used for every file: one for forward read and another for reverse. This primer does not appear in the contig as the contig was only formed as an assembled string of bases containing the basecalls of each sequence. A primer sticks to the DNA template, and its use initiates replication also necessary for DNA sequencing and Polymerase Chain Reaction (Lyons, 1998). The bases within the primer are not a part of the sequence. In order to incorporate the primer information in the correct sequence, the corresponding part from the original files is copied. However, while assessing the accuracy of the basecalls this part is not used. For accuracy calculations only the bases that are a part of the contig are considered.

The overall process of database preparation for accuracy measurement and comparison proved very complex and time-consuming. Overall, it involved obtaining the *ABI* basecalls, *Phred* basecalls, searching the large contig for correct sequences, re-writing to a new file the information from the contig where the sequence overlaps, and finally adding the primer to the new file. These steps finally resulted in a five-part database of *ABI* raw data files, a large contig of DNA bases assembled from that raw data, the *ABI* basecalls, the *Phred* basecalls, and the correct basecalls. Using the prepared database and different comparative techniques, the percent accuracy for *TraceTools* was found and compared with those of *ABI* and *Phred*. The outcome is presented in the results section.

2.2 DATA PROCESSING AND BASECALLING

The algorithm for basecalling used by *TraceTools* is based on processing the raw data contained in the *ABI* sequencing files. The general approach is oriented toward preserving the information contained in the raw data and avoiding the use of traditional filtering techniques. A detailed presentation of the proposed approach is presented in Domnisoru et. al., 2000 and Domnisoru and Musavi, 2003. The algorithm has two main steps: 1) **data processing** - where the raw data information is filtered and a model for the spacing between consecutive bases is constructed, and 2) **basecalling** - where the base spacing information is used to predict the location of the bases and make basecalls.

Several pre-processing steps are employed to ensure the extraction of a model for the base spacing from the raw data file. A preliminary filtering is applied to smooth the signals. The cross talk parameters are detected automatically from the trace information and the cross talk removal itself is applied to the variation of the signals as opposed to the signals directly. The signals are reconstructed from their variation and aligned at a baseline. The next step is the detection of the peak candidates based on the local “peakness” and height of the signals. Using the peak candidates, a preliminary model for the base spacing is determined. The peak candidates not fitting the model are eliminated followed by a recalculation of the base spacing model. Note that the base spacing model is differentiated for each combination of possible two consecutive bases. There are 16 such combinations one for each pair (AA, AC, AG, AT, CA, CC, CG, ...). Therefore, the base spacing model has 16 “sub-models” for each possibility.

The final step, the basecalling, is a hybrid algorithm based on the prediction of the spacing between bases and a fuzzy logic basecall. At every base, say A, the base position is submitted to the four models (AA, AC, AG, AT) that start with the current base and the location of the next four possible base candidates are predicted. The algorithm then evaluates the peakness of the signals, the height, and the slope for all the four predicted locations on a local basis. A fuzzy system would then

use the peakness, the height, and the slope information to make a basecall. After the basecalling is performed once, the base spacing model is recalculated and the basecalling part is redone using the updated spacing information.

2.3 CONFIDENCE VALUES

As the basecalling algorithm's error rates drop, the smaller basecall errors could be difficult to locate. Hence, assembling algorithms and human operators use a confidence value measure to determine how well the basecalling algorithm has performed for each basecall. This will clearly make it easier to uncover potential errors and correct them, thus increasing the throughput of genetic sequencing. So far, establishing such a confidence value for the basecalls was done primarily in support of the development of the *Phred* basecalling system (Ewing et. al, 1998). They provided a predictive measure, known as quality value, which would directly correlate to true trace error rates and help locate possible errors. This model utilized one static model for all sequences and it is based on sequencing error rate not the local information at each base, hence, resulting in a not so reliable measure of success for basecalls. Furthermore, their model would be inadequate with continuously evolving new sequencing techniques and process variability.

ABI software previously didn't provide any confidence value for its basecalls. However, its latest version, *ABI 3730*, provides a measure that is similar to *Phred* quality value and in fact it uses the same technique. The only difference is that *ABI 3730* has replaced *Phred's* numerical values with color coded bars for easy and quick recognition of error areas. In contrast, our model employs fuzzy logic, which provides flexibility, adaptability and intuition through the use of linguistic variables and fuzzy membership functions. More importantly, the measure is a true confidence values based on the information at the base.

In the fuzzy model, three trace features from the basecalling algorithm are collected and used as inputs (French et. al, 2002). The first feature is the *height*, which is simply the height of a peak. The

second is the *peakness*, which is a measure related to the concavity at the top of a peak. The final feature is the *base spacing*, which is the difference in location from one peak to another. In addition to the first most likely candidate (the base “called”), the *peakness* and *height* are also found for the second (“2nd”) likely candidate and used in the fuzzy model.

As shown in Figure 1, the fuzzy system involves four subsystems that are designated as *Fuzzy Peakness*, *Fuzzy Height*, *Fuzzy Spacing*, and *Fuzzy Confidence*. The first three subsystems calculate C_p , C_H , and $C_{\Delta S}$ based on peaknesses (P_{called} and $P_{2\text{nd}}$), heights (H_{called} and $H_{2\text{nd}}$), and spacings ($|\Delta S_{\text{previous}}|$ and $|\Delta S_{\text{next}}|$), respectively. The linguistic terms for C_H are defined as very low (VL), low (L), medium (M), high (H), and very high (VH) and for C_p and $C_{\Delta S}$ are defined as low (L), medium (M), and high (H). The *Fuzzy Confidence* system takes in the confidence values provided by the other three subsystems and computes the overall confidence value (C) of the base called. The fuzzy linguistic terms for C are very low (VL), low (L), medium (M), high (H), and very high (VH). Note that since there are 3 input variables for the fuzzy confidence subsystem, there could be as many as 45 ($3 \times 5 \times 3$) rules, of which some are unlikely to happen. The fuzzy operator AND is used for all fuzzy rule premises involved in the subsystems and the confidence value of *height* is given more weight in setting up the fuzzy rules.

Figure 1

2.4 *TRACETOOLS* SOFTWARE

The above algorithms for data preprocessing, basecalling, and confidence value have been integrated in a software, called *TraceTools*. This software is designed to process *ABI 3700* chromatograms. *TraceTools*, which is a Windows based software, can display both the raw and the processed data after making basecalls. The display of raw data allows the user to view the data as recorded by the sequencing machine. When the basecalls made are uncertain, this display feature would help the user make confident decisions after investigating the raw data. The basecalling 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 basecall through a color-coded rectangular bar. After calling bases, *TraceTools* can write the sequences to files in FASTA format. Below is a screenshot of *TraceTools* window showing both the raw data (upper window) and its processed version and called bases (lower window). 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.

TraceTools' confidence value is a continuous value between 0 and 1, with 1 indicating the highest confidence. 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 are 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. Figure 2 shows bases with very high confidence (full green), medium confidence (yellow) and poor confidence (red).

Figure 2

3. RESULTS AND DISCUSSION

In this section, the results of *TraceTools* testing is presented and compared with the *ABI* and *Phred* software. The accuracy of *TraceTool* basecalling for a total of 3362 chromatogram files, as described in the data preparation section, was recorded. 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 basecalling 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 basecalls by *ABI*, *Phred*, and *TraceTools* were compared against the correct sequences. The overall accuracy for *TraceTools* was higher than *ABI* and *Phred*. 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. From Figure 4, it can also be seen that *TraceTools* had 100 more files than *Phred* and 300 more files than *ABI* in the high accuracy range of 95-100%. In other words *TraceTools* had 3% more files than *Phred* and 9% more files than *ABI* that were highly accurate.

Our experiments with *TraceTools* confidence values also provided more accurate representation of the basecall 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 ease of recognition. 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 basecaller 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 basecall 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

Figure 4

Figure 5

Table 1

There are a couple of points that one needs to consider regarding the accuracy results presented above. One point is the fact that the contig is not 100% full proof. In other words, the contig that was used for the extraction of correct sequences has shown to have some errors in it. And the other point is that the accuracy results may be biased in the favor of *Phred* because as it was explained before, *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 *TraceTools* 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 base at the vertical line 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 basecall. This fact is also verified by *ABI* and *Phred*, as indicated on the figure. Even though *TraceTools* makes a correct basecall 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 basecaller results and the contig. As evident from the results *TraceTools* makes the basecall *C* similar to the contig, but *ABI* calls it a *G* and *Phred* calls it a *T*.

Figure 6

Figure 7

4. CONCLUSION

A new sequencing system, *TraceTools*, has been presented that has shown improved results over the *ABI* and *Phred* software on both the called bases and their confidence values. The increase in the accuracy can be attributed to *TraceTools* unique features of filtering, mobility shift correction, adaptive base space prediction, cross talk removal, and fuzzy confidence value calculation. The color-coded confidence values assist the user in quick identification of bases with low confidence values.

AVAILABILITY

A copy of *TraceTools* trial version for sequencing ABI 3700 files can be downloaded from the project web site at: <http://www.intsys.maine.edu/AccurateDNA.htm#Downloads>.

ACKNOWLEDGEMENTS

This research was funded in part by NSF grant #DBI-0090738 and #EEC-9820332. The authors would like to acknowledge Dr. Ralph Dean of the Fungal Genomics Laboratory at North Carolina State University for generously providing the *ABI* raw data and corresponding contig used for this study. The authors also extend their appreciation to Ms. Ceara McNally and Mr. Brian French for software development and database preparation.

REFERENCES

- ABI PRISM 3700 DNA Analyzer Specifications*, at: http://www.appliedbiosystems.com/products/specifications.cfm?prod_id=40.
- Domnisoru, C., Zhan, X., and Musavi, M.T., (2000), *Electrophoresis*, **14**, 2983-2989.
- Domnisoru, C. and Musavi, M., (2003), Method for reducing cross talk within DNA data, United States Patent Office #6,598,013.
- Ewing, B., Hillier, L., Wendl, M. C., and Green, P., (1998), *Genome Research*, **8**, 175-185.
- Felsenfeld, A., Peterson, J., Schloss, J., and Guyer, M., (1999), *Genome Research*, **9**, 1-4.
- French, B., Domnisoru, C., Resson, H., and Musavi, M.T., (2002), Proceedings of the International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS), Las Vegas, Nevada, USA, June 27, 2002, 203-209.
- Koonin, S., (1997), Genome Informatics, Human Genome Project, The MITRE Corporation, JASON Program Office.
- Lyons, R. H., (1998), A Molecular Biology Glossary
<http://seqcore.brcf.med.umich.edu/doc/educ/dnapr/mbglossary/mbgloss.html>.
- McNally, C., Domnisoru, C., and Musavi, M.T., (2002), Proceedings of the International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS), Las Vegas, Nevada, June 27, 2002, 217-223.
- Olson, M. and Green, P., (1998), *Genome Research*, **8**, 414-415.
- Richerich, P., (1998), *Genome Research*, **8**, 251-256.
- Tatusova, T.A. and Madden, T.L., (1999), *FEMS Microbiol Lett.*, 174:247-250.

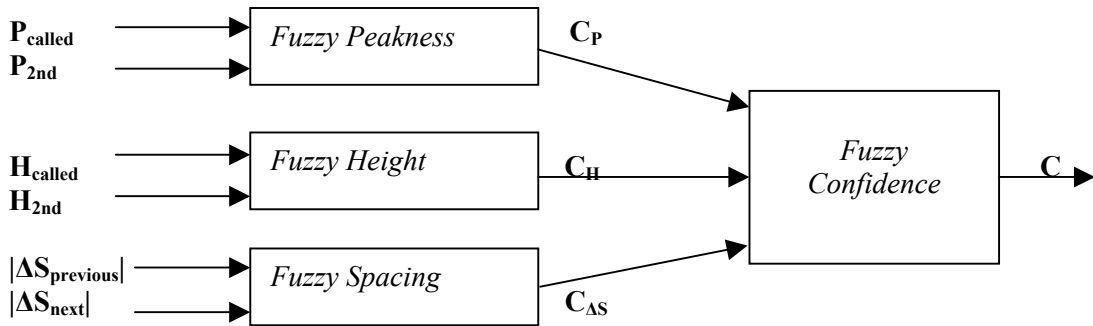


Figure 1: Block diagram of the overall fuzzy logic system for calculation of confidence value.

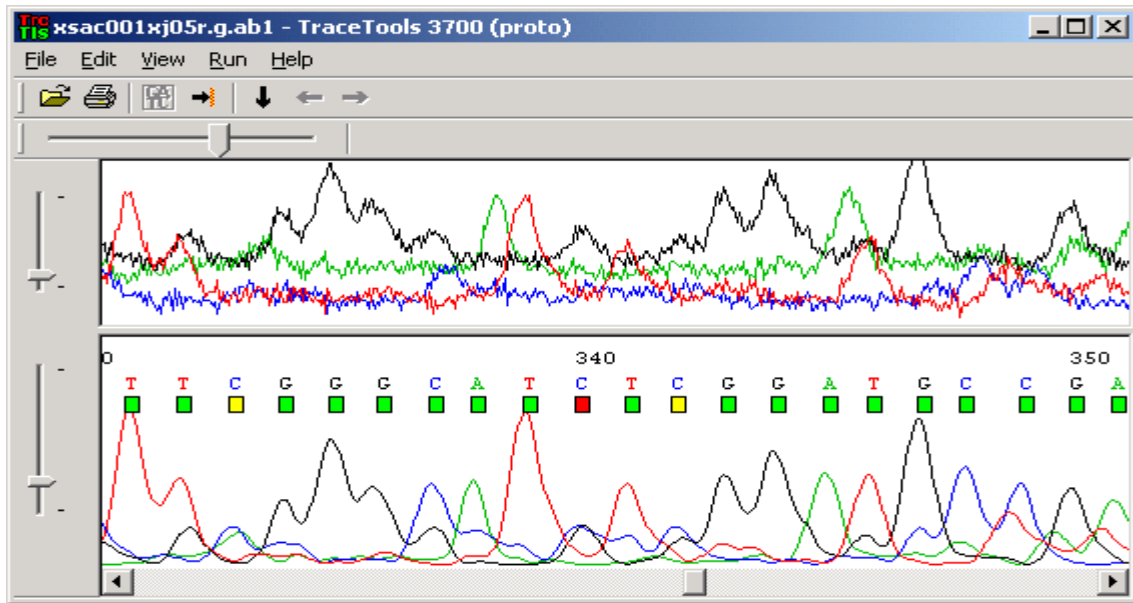


Figure 2: *TraceTools* window showing a sequence with bases called and their confidence values indicated by color-coded bars; upper window shows unprocessed data.

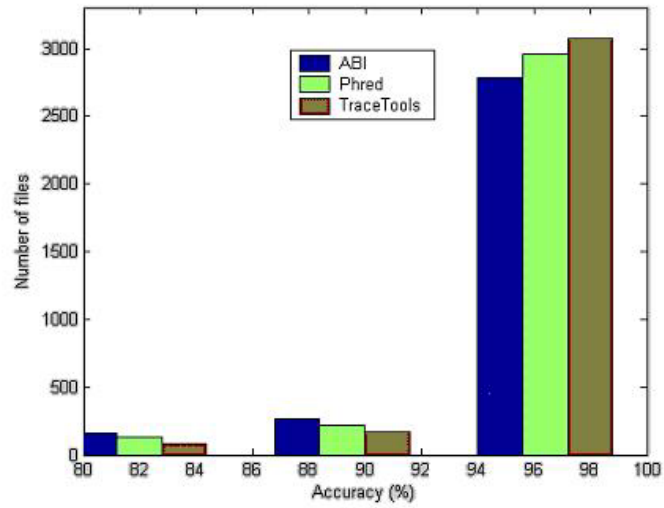


Figure 3: Histogram of accuracy for *ABI*, *Phred* and *TraceTools* for range 80-100%.

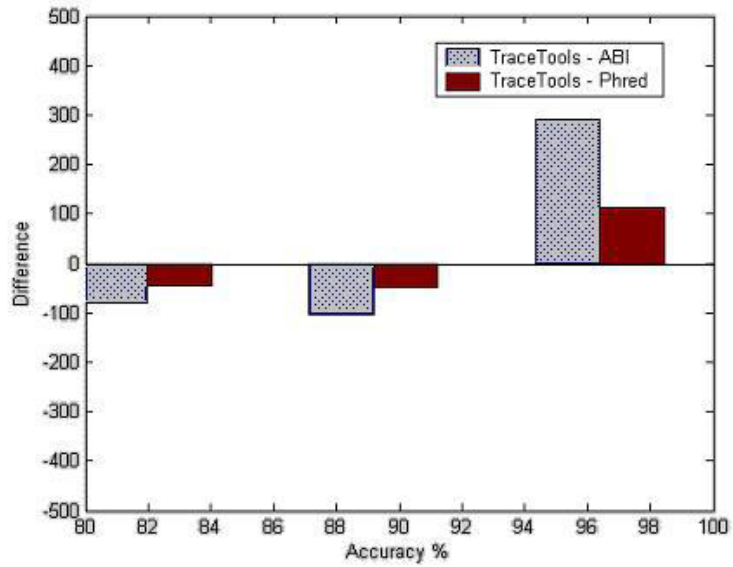


Figure 4: Difference histograms between *TraceTools* and *ABI* & *TraceTools* and *Phred*.

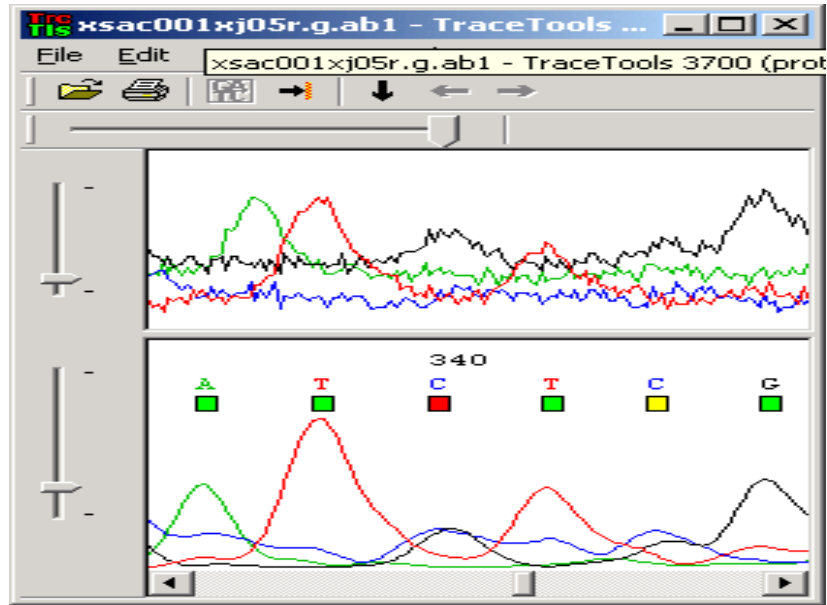


Figure 5: A screenshot of *TraceTools* for confidence value analysis.

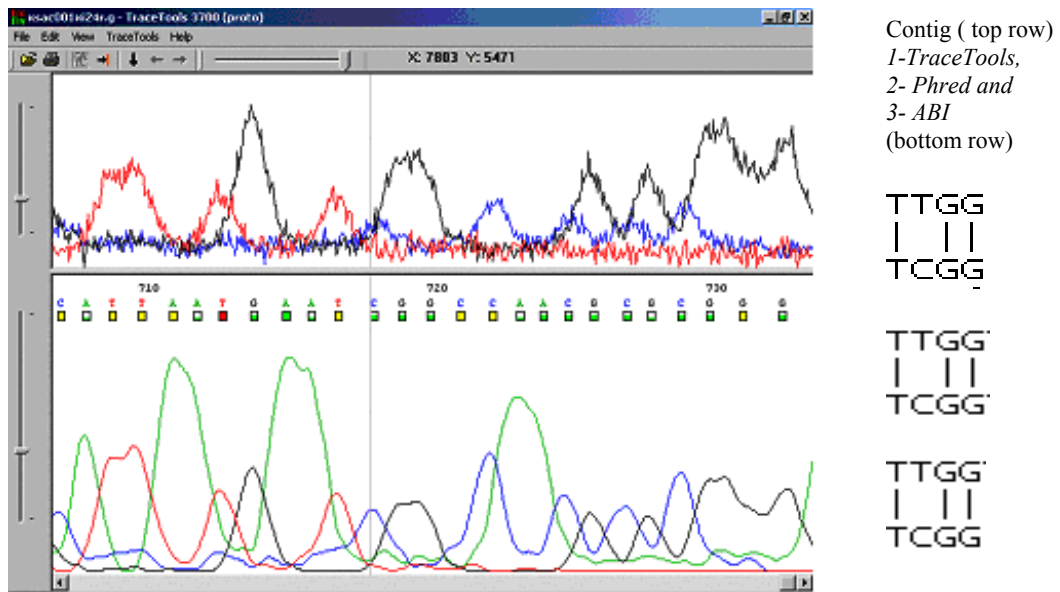


Figure 6: Screenshot of *TraceTools* results for a chromatogram file (upper window: raw data; lower window: base calls made by *TraceTools*) (Right) *TraceTools*, *Phred* and *ABI* base calls w.r.t. Contig; vertical lines show match between them.

Contig → T T T C C T Contig → T T T C C T Contig → T T T C C T
 Phred → | | | | | ABI → | | | | | TraceTools → | | | | |
 T T T T C T T T T G C T T T T C C T

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

	Confidence values for the bases called					
	A	T	C	T	C	G
<i>Phred</i> (Max value 50)	20	12	14	13	15	22
<i>TraceTools</i> (Max value 1)	0.84	0.92	0.12	0.78	0.22	0.90

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