2004

Confidence Measure for DNA Base Calling Using a Fuzzy System

Rency Susan Varghese

CONFIDENCE MEASURE FOR DNA BASE CALLING USING A FUZZY SYSTEM

By

Rency Susan Varghese

B.Tech. Kerala University (India) 1999

A THESIS

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

(in Electrical Engineering)

The Graduate School

The University of Maine

May, 2004

Advisory Committee:

Mohamad T. Musavi, Professor of Electrical and Computer Engineering, Co-Advisor

Habtom Ressom, Assistant Professor of Electrical and Computer Engineering, Co-Advisor

Bruce Segee, Associate Professor of Electrical and Computer Engineering

CONFIDENCE MEASURE FOR DNA BASE CALLING USING A FUZZY SYSTEM

By Rency Susan Varghese

Thesis Co Advisors: Dr. Mohamad T. Musavi
Dr. Habtom Ressom

An Abstract of the Thesis Presented
in Partial Fulfillment of the Requirements for the
Degree of Master of Science
(in Electrical Engineering)
May, 2004

Base calling is the central part of any large-scale genomic sequencing effort. Current sequencing technology produces error rates less than 3.5%. This corresponds to at least 35 errors in a 1000 base read. As the base calling algorithm's error rates drop, the smaller base call errors could be difficult to locate. Hence, assembling algorithms and human operators use a confidence value measure to determine how well the base calling algorithm has performed for each base call. This will clearly make it easier to uncover potential errors and correct them, thus increasing the throughput of genetic sequencing. The model developed here employs fuzzy logic, providing flexibility, adaptability and intuition through the use of linguistic variables and fuzzy membership functions. The proposed approach uses a fuzzy logic system to provide the confidence values of bases called. Three variables that

are calculated during the base calling procedure are involved in the fuzzy system. These variables can be calculated at any spatial location and are: *peakness*, *height*, and *base spacing*. In addition to the first most likely candidate (the base called), the *peakness* and *height* are also found for the second likely candidate. The technique has been tested on over 3000 ABI 3700 DNA files and the result has shown improved performance over the existing *Phred's* and *ABI's quality value*.

ACKNOWLEDGEMENTS

I would like to thank all the members of my committee: Dr. Mohamad T. Musavi, my advisor, for all the help and providing me with the opportunity to pursue my Master's degree at University of Maine; Dr. Habtom Ressom, my co advisor for his aid and advise for publishing papers; and Dr. Bruce Segee, who first introduced me to fuzzy logic. I want to thank the other entire faculty and staff members in the department and Intelligent Systems Lab who have given me help during my graduate study at University of Maine.

I wish to thank Padma and Mike for helping me with my thesis work. Also thanks to Siva, Kamal, Driss and Wayne for being nice friends and helping me with everything around Intelligent Systems Lab.

Finally, I would like to thank my husband, Anish Senan and my family for their support and care, without which this thesis wouldn't have been possible. I also extend my thanks to all my friends for their encouragement and assistance.

TABLE OF CONTENTS

1 Introduction

1.1 Purpose of Research

The unspoken goal of research into base calling algorithms is to attain 100% accuracy, thus eliminating the need for any intervention to determine the correct sequence. But given the current state of the art, more pragmatic goals for the next few years are for error rates around 1%. Although this is very low, it is certainly not zero, meaning that intervention, including consensus algorithms [1] and human operators, cannot be eliminated anytime soon. Paradoxically, as base calling algorithms' error rates drop, the smaller base call errors can become obfuscated and difficult to locate. That is why assembling algorithms and human operators use the confidence value measure to determine how well the base calling algorithm has performed at particular base calls, making it easier to uncover potential errors and to correct them, thus increasing throughput of genetic sequencing. It is unmistakable that confidence value prediction has emerged as an essential tool in contemporary genome mapping projects.

1.2 Objective

The objective of this thesis is to develop a novel algorithm that can predict the confidence values for each base called in DNA sequencing.

The proposed approach uses a two-stage fuzzy logic system to provide the confidence values of bases called. The algorithm developed can be

integrated with any DNA sequencing software. It can also be used as a measure to improve the accuracy of the DNA base caller.

1.3 Previous Work

By far the main body of work accomplished in the area of confidence value was done primarily in support of the development of the *Phred* base calling system [2]. *Phred's* work produces a predictive quality value measure that would directly correlate to true trace error rates. This value is used in discriminating where possible errors are located. By employing an algorithm [2] on a large data set *Phred* was able to create a model (a lookup table). The input space of the model consists of trace data features like *peak spacing, uncalled/called ratio, and peak resolution*. The output space is the resulting quality value, which should relate to the error probability of a base call by the following equation:

$$q = - 10 \bullet \log_{10}(e)$$

where q is the quality value and e is the error probability. Thus a base call having a probability of 1/1000 of being incorrect is assigned a quality value of 30. The error value was *log* transformed because the error probabilities *Phred* was working with were small.

One contributing measure that *Phred* introduced was the discrimination power of *quality value*. That is, how well the system locates the regions with errors and the regions that are error free. For example, if there is a base call sequence that contains 5 errors within a 100 base trace, a perfectly correct *quality value* for each base call could be the value 13. This

number comes from the fact that each base call is given an error probability of 5/100. So the confidence value is calculated by $(-10) \cdot \log_{10}(5/100) \approx 13$. Even though the *quality value* is correlated to the error rate it doesn't give us any idea where the errors are located. A better example would be splitting the 100 bases in half into two groups of 50 bases each. Suppose also that we find the first half has 4 errors and the second half has 1 error. This would mean that the bases in the first half could all be assigned the error probabilities 4/50, while the other half of the bases could be assigned 1/50, thus corresponding to confidence values of 11 and 17 respectively. We see that this example does a better job at discriminating the poor region (the first half) from the region that performed well (the last half). This leads to *Phred's* definition of discriminating power at the error rate:

$$P_r = \frac{|B_r|}{|B|}$$

where P_r is discriminating power factor for error rate r. $|B|$ is the number of bases in set B and $|B_r|$ is the number of bases in B_r. P_r measures the effectiveness of the error probability assignments at extracting a subset of bases having a lower error rate r.

Though this method has gained wide acceptance, employing just one lookup table for all sequences leads to an inflexible model. As sequencing machines, sequencing chemistry, and base calling algorithms improve; models must adapt in order to reflect the technological progress. Even worse there can be variations between sequencing machines that can compromise the model rendering it not truly predictive of the error. Also this system does

not allow for the model to adapt to newer base calling techniques, variations in sequencing machines, and deviations in other quality control measures. All of this leads to an inflexible model that doesn't forward any intuition with the trace features as they relate to the confidence value.

1.4 The Proposed Approach

The proposed approach uses a two-stage fuzzy logic system to provide the confidence values of bases called. As opposed to *Phred's quality value,* this method uses three variables that are calculated during the base calling procedure. These variables can be calculated at any spatial location and are: *peakness, height,* and *base spacing*. In addition to the first most likely candidate (the base called), the *peakness* and *height* are also found for the second likely candidate. The three sets of variables are then fed into three separate fuzzy sub systems and confidence values corresponding to *height, peakness* and *base spacing* are calculated. In the second stage, another fuzzy sub system takes in the confidence values provided by the other three subsystems and computes the overall confidence value of the base called. The results of this research have shown improvement over the quality values provided by *Phred*.

1.5 Why Use Fuzzy Logic

Fuzzy Logic is a paradigm for an alternative design methodology that can be applied in developing both linear and non-linear systems. Fuzzy logic lets one use human knowledge and experience to describe complex systems using simple English-like rules. It does not require any system modeling or complex mathematical equations governing the relationship between inputs and outputs. It typically takes relatively few rules to describe systems that may require numerous lines of conventional software. As a result, Fuzzy Logic often significantly simplifies design complexity. With fuzzy logic design methodology some time consuming steps are eliminated. Moreover, during debugging and tuning, one can easily change the system by simply modifying rules, rather than redesigning the whole system. In addition, since fuzzy logic is rule based, one can focus more on the application instead of programming.

For computing the confidence values of the bases called by a DNA base caller, fuzzy logic helps to incorporate the information collected from the operators/users in a simple way. Debugging can be easily performed using the information from the operators.

1.6 Thesis Organization

This thesis is divided into seven chapters. Chapter 2 gives an introduction on DNA sequencing, data preprocessing, base calling and database preparation. Chapter 3 discusses the ideas behind the proposed thesis and discusses in detail the input data extraction for the model to be

developed. Chapter 4 explains the implementation of the confidence fuzzy model and describes the fuzzy rules and membership functions used for the development of the model. Chapter 5 illustrates the analyses and results of the confidence fuzzy model. It also shows a comparison study on the confidence values with the *Phred's 'quality values'*. Chapter 6 discusses on how the fuzzy confidence system can be used to improve the accuracy of DNA basecalling. Finally, Chapter 7 concludes the discussion of the topic and proposes future work on the method.

2 DNA Sequencing and Database Preparation

The proposed technique for calculating confidence value has been integrated in a novel base calling software called *TraceTools*, developed at The Intelligent Systems Laboratory, University of Maine. This chapter provides background information about DNA sequencing, *Tracetools* and about the database prepared for testing the software developed.

2.1 DNA Sequencing

The technology for sequencing DNA has rapidly evolved from gel based to capillary electrophoresis (CE) [3]. The most widely used sequencing systems are the *ABI* (Applied Biosystems Inc.) sequencing machines [4]. 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. The output of this physical process is affected by noise, but the interference between the four filters and other phenomena is less understood.

Although the sequencing machines have evolved, 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). Since the peaks obtained will not be clearly

separated and not big enough when compared to the noise at the baseline, automated sequencing software are needed to find the peaks and make an accurate base calling for the data. By far, the ABI software and *Phred* have dominated the sequencing community.

2.2 *TraceTools*

TraceTools is base calling software [5] that utilizes the fuzzy confidence value model developed in this thesis. For developing and testing the *TraceTools*, a comprehensive database of *correct* DNA sequences corresponding to the ABI raw data was constructed. This comprehensive database was used for comparing the accuracy of *TraceTools* with other popular base calling programs such as *Phred* and *ABI*.

2.3 Database Preparation

The database preparation involves creating a database of sequences each corresponding to a raw data ABI file to evaluate the performance of base calling programs [6]. These sequences must contain the correct bases so that they can be used as the ground truth for comparing it with the results obtained by base calling programs. To accomplish this, a contig, a 300,000 base long sequence comprised of thousands of overlapping sequences is used. The accuracy of these ground truth sequences necessarily depends on the accuracy of the contig. The contig and *ABI* raw data was obtained from

the North Carolina State University. The raw data was generated by the *ABI* 3700 system.

2.4 Data Processing and Base Calling

The algorithm for base calling 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 approach is presented in [7] and [8]. While *Phred* uses ABI's preprocessed data, *Tracetools* starts with the raw data. The algorithm has two steps: 1) **data processing -** where the raw data information is filtered, color separated and a model for the spacing between consecutive bases is constructed, and 2) **base calling** - where the base spacing information is used to predict the location of the bases and make base calls.

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 and the cross talk removal itself is applied to the variation of the signals (as opposed to the tradition of using 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. A preliminary model for the base spacing is determined, and the peak candidates from the good region 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 and therefore, the model has 16 "sub-models" for each possibility.

The final step, the base calling, is based on the prediction for the spacing between bases. The base call, evaluates the *peakness* of the signals, the *height*, and the slope on a local basis. After the base calling is performed once, the base spacing model is recalculated and the basecalling part is redone using the updated spacing information. After the basecalling is done, the same variables, *peakness, height* and *base spacing* are used to find the confidence values of the bases being called.

The results of the comparison of accuracy of *TraceTools* with other popular base calling programs show an average accuracy of 97.28% for *TraceTools*, 97.10% for *Phred,* and 95.99% for *ABI*.

3 Data Extraction for Confidence Value Calculation

This chapter describes the raw data and the input data extracted for the calculation of confidence values.

3.1 Raw Data

The ABI system for DNA sequencing collects four signals corresponding to the four bases C, T, A and G, as shown in Figure 3-1. The measured signals represent fluorescence intensities at four different wavelengths.

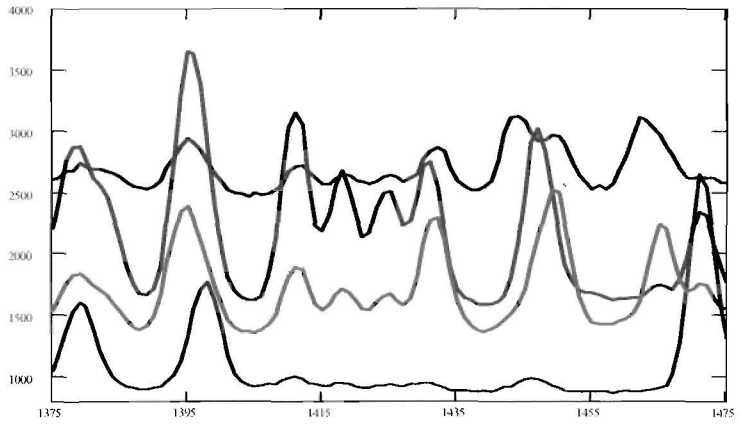

Figure 3-1 Raw data collected from ABI machines.

The raw data, captured by the sequencing machine is first filtered and prepared as a succession of peaks. The stream of peaks is then processed and basically, each peak is associated with a base.

3.2 Input Data Extraction for Confidence Calculation

The initial motivation for developing the confidence model was so that the basecalling algorithm could have a confidence value to check the performance of the system. Trace features are collected from the raw data and are used as inputs to the fuzzy model. These are the key parameters that help in identifying the bases correctly and also predict the confidence values. They appear to play a role in intuitive human assessments of confidence values. In this fuzzy model, three trace features are collected from the basecalling algorithm. The first feature is the *height* (*H*), i.e., the height of the peak as in Figure 3-2. The second is the *peakness* (*P*), which is a measure related to the concavity at the top of a peak Figure 3-3. The final feature is the *base spacing* (ΔS), i.e., the location differences from one peak to another.

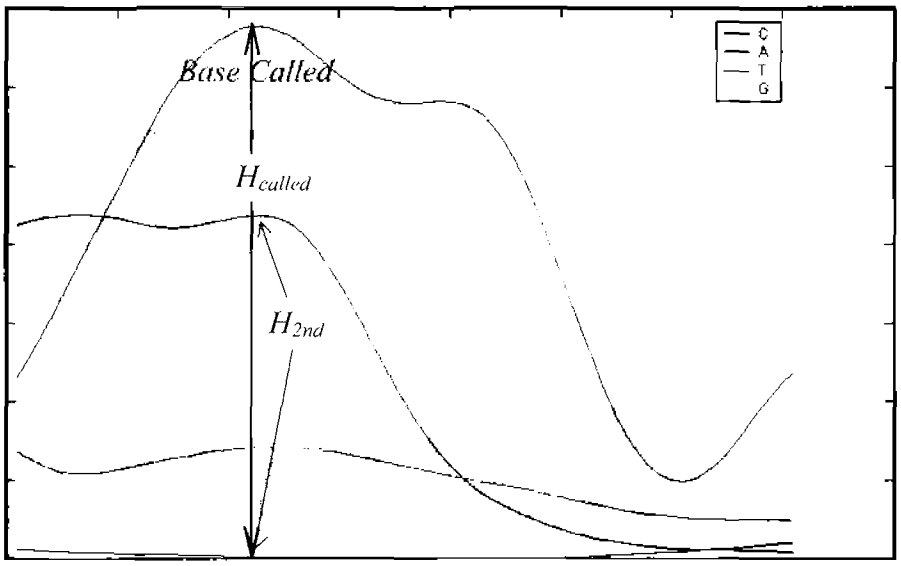

Figure 3-2 Representation of *height* variables

In addition, the base calling algorithm not only identifies the most likely base call candidate within a local position, but also the second most likely base call candidate. This gives a starting point from which we can define input variables to the fuzzy system. The input variables are explained in detail below:

Height: *Height* is calculated as the amplitude of each base from the baseline.

H_{called}: Height of the base called.

H_{2nd} : Height of the 2^{nd} candidate.

Peakness: *Peakness* is an indication of how sharp a peak is locally. It is defined for the entire trace, not just where a peak is located. Therefore, the higher the *peakness*, there is a greater chance to have a peak in that location. The mathematical calculation for *peakness* is described in the next subsection.

P_{called} : *Peakness* of the base called.

P_{2nd} : *Peakness* of the 2^{nd} candidate.

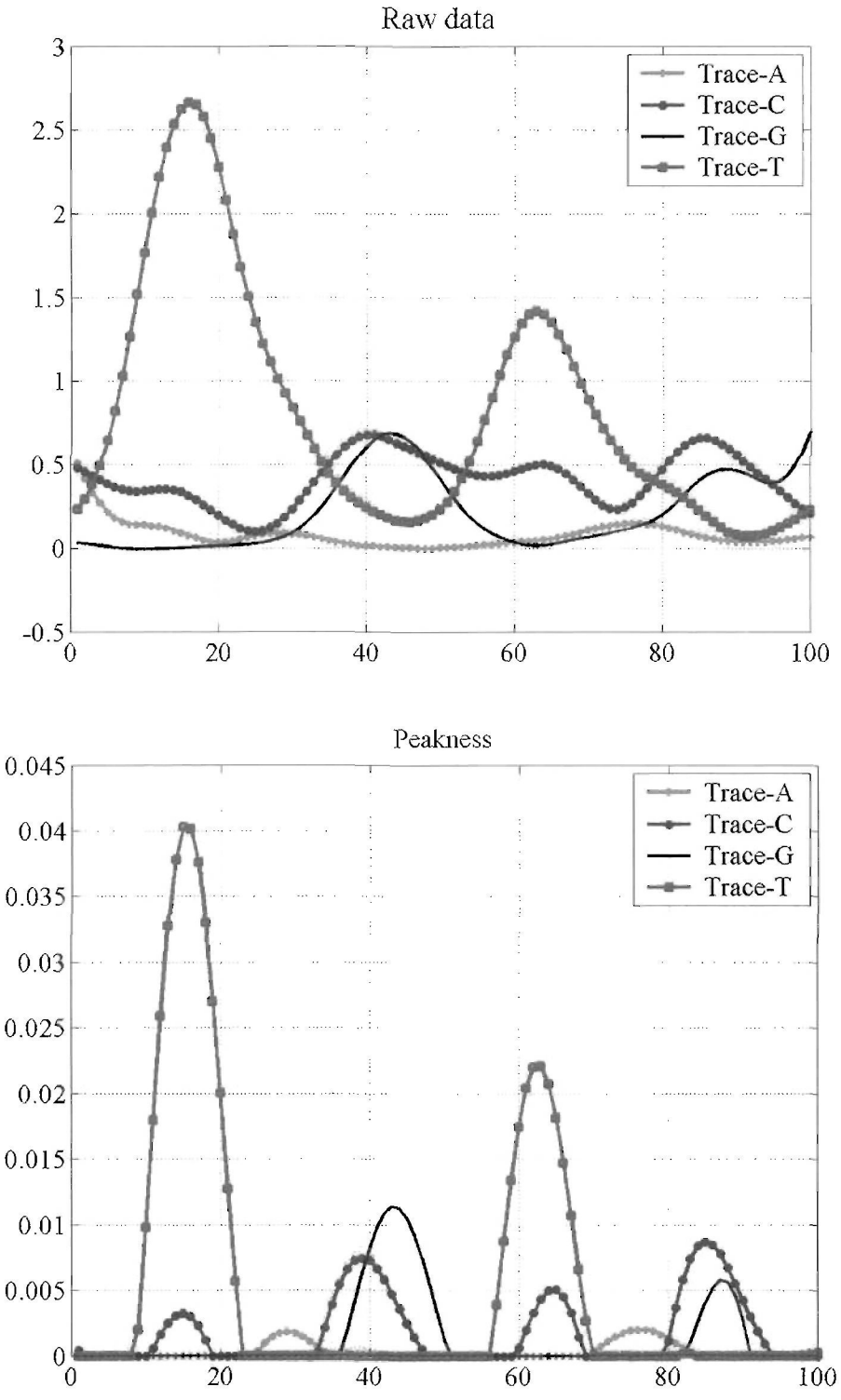

Figure 3-3 Representation of *Peakness* (below) and Raw data (top)

3.3 *Peakness* Calculation

As shown in Figure 3-4, circle of curvature is a circle that "fit" the curve at a point. If the curve is turning sharply, the radius of curvature is small and if the curve is turning slowly, the radius of curvature is large. Therefore *peakness* can be calculated as the inverse of the radius of the largest circle that could be drawn to be tangent at the curve.

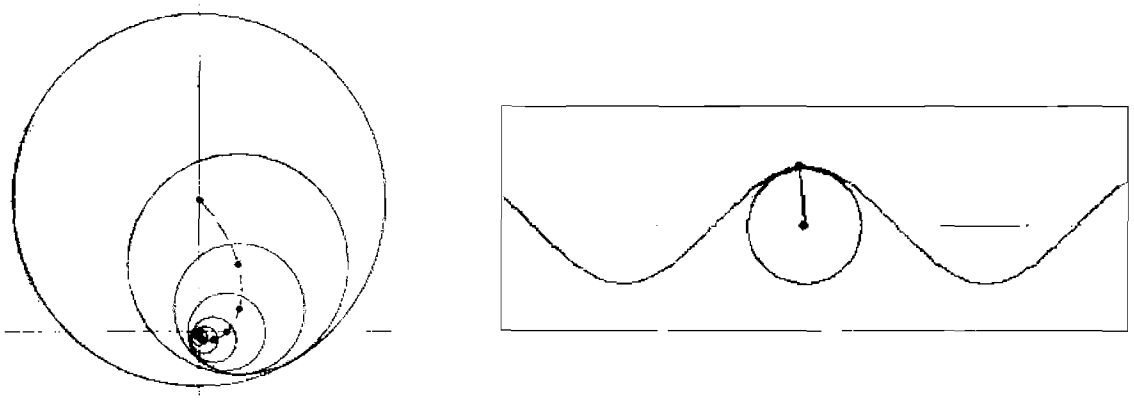

Figure 3-4 Osculating Circle and Radius of Curvature

The radius of curvature is given by

$$R = \frac{1}{k}$$

where k is the curvature. At a given point on a curve, R is the radius of the osculating circle (*The circle that shares the same tangent as a curve at a given point*). Let *x* and *y* be given parametrically by

$$x = x(t)$$
$$y = y(t)$$

Then the curvature k is defined by [9,10]

$$k = \frac{d\phi}{ds} = \frac{\frac{d\phi}{dt}}{\frac{ds}{dt}} = \frac{\frac{d\phi}{dt}}{\sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}} = \frac{\frac{d\phi}{dt}}{\sqrt{x'^2 + y'^2}}, \tag{3.1}$$

where ϕ is the tangential angle and s is the arc length. As can readily be seen from the definition, curvature therefore has units of inverse distance.

$\frac{d\phi}{dt}$ in the above equation can be found using the identity

$$\tan\phi = \frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{y'}{x'}, \tag{3.2}$$

so

$$\frac{d}{dt}(\tan\phi) = \sec^2\phi\frac{d\phi}{dt} = \frac{x'y'' - y'x''}{x'^2} \tag{3.3}$$

and

$$\frac{d\phi}{dt} = \frac{1}{\sec^2\phi}\frac{d}{dt}(\tan\phi) = \frac{1}{1+\tan^2\phi}\frac{x'y'' - y'x''}{x'^2}$$

$$= \frac{1}{1+\frac{y'^2}{x'^2}}\frac{x'y'' - y'x''}{x'^2} = \frac{x'y'' - y'x''}{x'^2 + y'^2} \tag{3.4}$$

Combining (1), (2) and (4) we get

$$k = \frac{x'y'' - y'x''}{(x'^2 + y'^2)^{3/2}} \tag{3.5}$$

For a two-dimensional curve written in the form $y = f(x)$, then the equation of curvature becomes

$$k = \frac{\frac{d^2y}{dx^2}}{\left[1+\left(\frac{dy}{dx}\right)^2\right]^{3/2}}. \tag{3.6}$$

Equation 3.6 is used in the fuzzy confidence algorithm for the calculation of *peakness* from raw data.

Spacing: Ideally, the spacing between two bases should be equal regardless of the base location. However, this is not observed in real DNA data due to the interaction between the dinucleotide sequences [11]. In order to account for the variation of the spacing, several approaches are possible. *Phred*'s [12] approach is to determine regions of equally spaced bases and analyze the sequence region by region. Giddings et al. [13] uses the space between bases by approximating the spacing with a polynomial and hence use the approximated value as input in a latter base call module. Although these are useful techniques, important information related to space between bases is already lost through the pre-processing steps.

Dominisoru and Musavi [14] created a base spacing model that has the spacing between each pair of possible bases. For example, the spacing variation between the bases A and G in this order can be significantly different than the spacing between G and A. According to this model, there are 16 different datasets corresponding to the possible combination of 2 bases. In this thesis, the above model is used to calculate the predicted

distance to the next base. Then ΔS_{next} is calculated as the difference between the *actual* (distance calculated between the called base and the next called base) and the *predicted* distance as illustrated in Figure 3-5.

$$\Delta S_{next} = S_{n\_actual} - S_{n\_predicted}$$

Similarly, for $\Delta S_{previous}$, the predicted distance to the previous base is obtained from the model and the difference between the *actual* (distance calculated between the called base and the previous called base) and the *predicted* is calculated. This becomes the input to the fuzzy model explained in the following chapter.

$$\Delta S_{previous} = S_{p\_actual} - S_{p\_predicted}$$

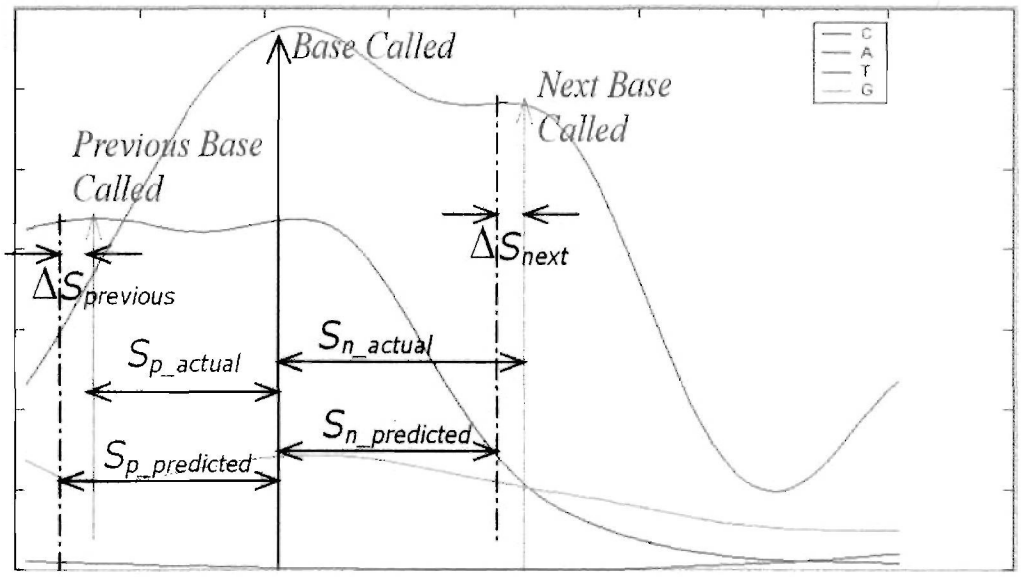

Figure 3-5 Representation of *spacing* variables

4 Confidence Fuzzy Model

This chapter describes the fuzzy model and the confidence value calculations.

4.1 Fuzzy Model

As shown in Figure 4-1, the fuzzy system involves four subsystems that are designated as *Fuzzy Peakness*, *Fuzzy Height*, *Fuzzy Spacing*, and *Fuzzy Confidence* [15]. The first three subsystems calculate C_P, C_H, and $C_{\Delta S}$ based on *peakness* (P_{called} and P_{2nd}), height (H_{called} and H_{2nd}), and spacing ($|\Delta S_{previous}|$ and $|\Delta S_{next}|$), respectively. The *Fuzzy Confidence* system takes in the confidence value provided by the other three subsystems and computes the overall confidence value (C_O) of the base being called.

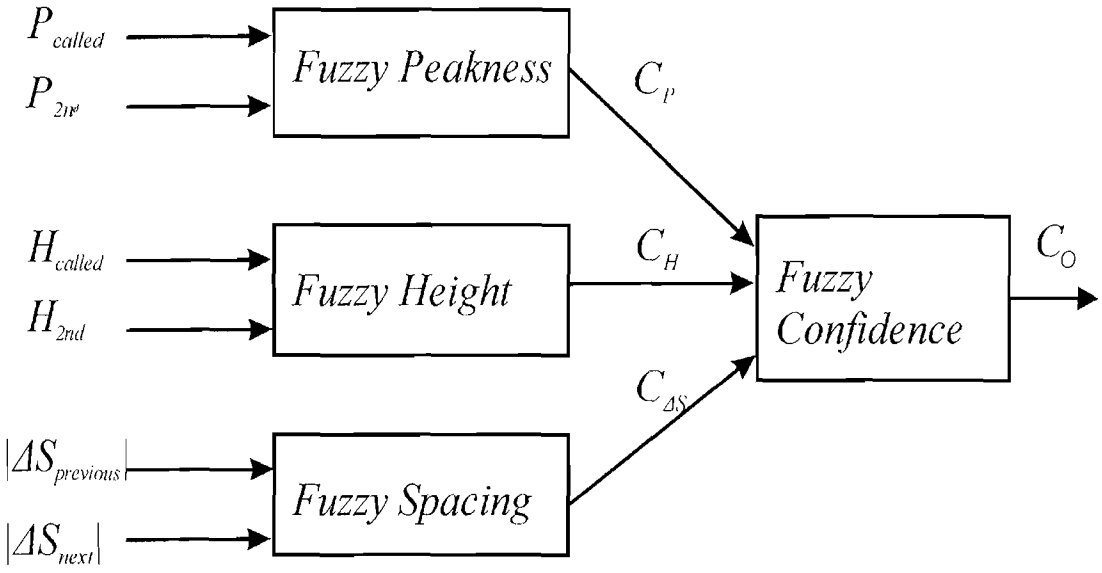

Figure 4-1 Block diagram of the overall fuzzy logic system.

The variables used in the *Fuzzy confidence* system are as below:

C_P: Confidence value of the base called relative to the *peakness (P)* variable.

C_H: Confidence value of the base called relative to the *height (H)* variable.

$C_{\Delta S}$: Confidence value of the base called relative to *base spacing* (ΔS) variable.

C_o: Overall confidence value of the base called.

4.2 Fuzzy Logic

The incentive for using fuzzy logic is so that we can take advantage of the linguistic variables feature inherent in fuzzy logic. It is also a natural extension of traditional Boolean Logic. To illustrate this point we should first entertain what is meant by traditional set membership with respect to Boolean Logic. In this case a value either has membership or does not have membership within a defined set. Instead of a value having a membership of 0 or 1, the degree of membership in Fuzzy Logic lies between 0 and 1 inclusively, allowing, for example, a value of 0.5 as a possible value.

To describe this, we can sample a certain population of people on whether or not it is warm outside over a varying degrees of temperatures and plot the number of people who think it is warm outside over a varying degrees of temperature and plot the number of people who think it is warm versus temperature. The result would be membership for the degree of truth

for 'warm' as seen in Figure 4-2, thus reflecting the naturally ambiguous term 'warm'

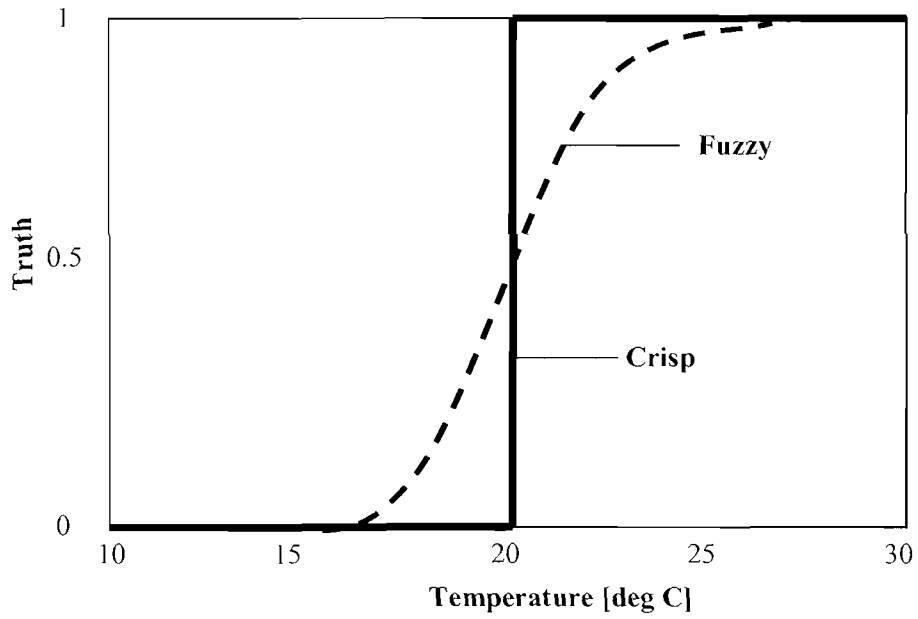

Figure 4-2 Fuzzy set of being Warm

Instead of having a membership of 0 or 1, the degree of membership in Fuzzy Logic lies between 0 and 1 inclusively. We can see that the membership function in Figure 4-2 captures the essences of what the linguistic term 'warm' means much better then two-valued logic ever could.

A fuzzy set is thus defined by a function that maps objects in a domain of concern to their membership value in the set [16]. Such a function is called a *membership function*. Also, the domain of membership functions is called the *universe of discourse.*

The next thing to consider is to choose the membership functions. What function should the fuzzy sets take, and how many regions should each universe of discourse be divided up into? There is no unique solution. We considered a trapezoidal membership function for our fuzzy models. For example, the trapezoidal membership function for *peakness* is depicted in Figure 4-3. Each fuzzy variable is then arbitrarily divided into 3 or 4 fuzzy sets, based on intuition.

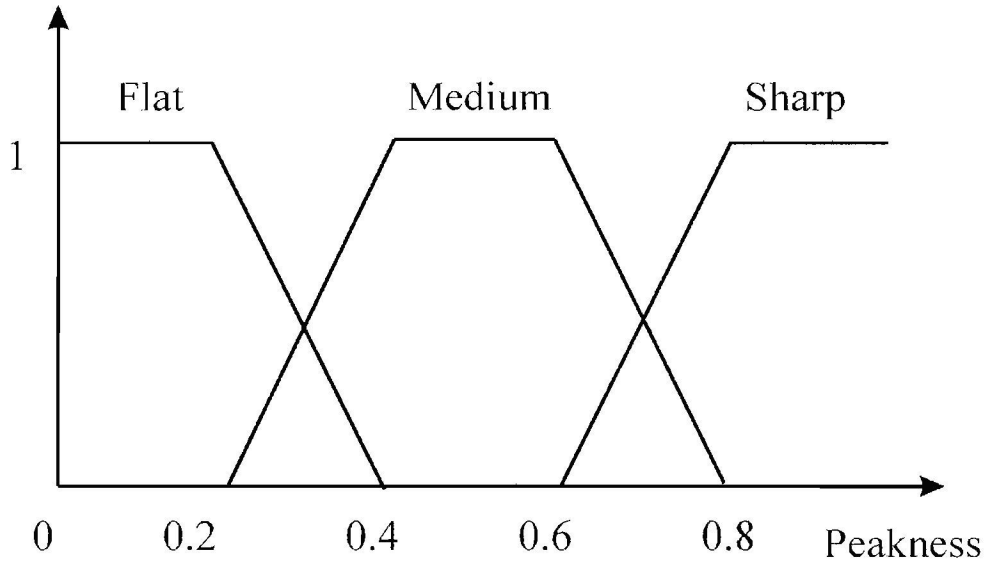

Figure 4-3 Trapezoidal Membership Function for *Peakness*

4.3 IF-Then Rules

The Fuzzy model used employed implications in the form of if-then rules. The fuzzy if-then rules are gleaned from intuition and experience.

For the Fuzzy Height sub system, each variable is divided into 5 regions of 'Very Low', 'Low', 'Medium', 'High', and 'Very High'. The linguistic terms for C_H are also defined as Very Low (VL), Low (L), Medium (M), High (H), and Very High (VH). Table 4-1 provides the fuzzy rules for this system. For example, the first cell of Table 4-1 indicates that: if the *height of* H_{called} is *'Very Low'* and H_{2nd} is *'Very Low'*, then the confidence value of *Fuzzy Height* subsystem is *'Low' (L)*.

Table 4-1 If-then rules for *Fuzzy Height* Subsystem

	H_{2nd}					
H_{called}		Very Low	Low	Medium	High	Very high
	Very Low	L	VL	VL	VL	VL
	Low	L	VL	VL	VL	VL
	Medium	M	L	VL	VL	VL
	High	H	M	L	VL	VL
	Very High	VH	VH	H	L	VL

For the *Fuzzy Peakness* subsystem, each variable is divided into 3 regions of 'Flat', 'Medium', and 'Sharp'. The linguistic terms for C_P are defined as Low (L), Medium (M), and High (H). Table 4-2 provides the fuzzy rules for this system.

Table 4-2 If-then rules for *Fuzzy Peakness* Subsystem

		P_{2nd}		
		Flat	Medium	Sharp
P_{called}	Flat	L	L	L
	Medium	H	M	L
	Sharp	H	H	M

Similarly, for the *Fuzzy Spacing* subsystem, each variable is divided into 3 regions of 'Small', 'Medium', and 'Large'. The linguistic terms for $C_{\Delta S}$ are defined as Low (L), Medium (M), and High (H). Table 4-3 provides the fuzzy rules for this system.

Table 4-3 If-then rules for *Fuzzy Spacing* Subsystem

		$\Delta S_{previous}$		
		Small	Medium	Large
ΔS_{next}	Small	H	H	M
	Medium	H	M	L
	Large	M	L	L

The fuzzy linguistic terms for the Overall Fuzzy Confidence System, C_o are Very Low (VL), Low (L), Medium (M), High (H), and Very High (VH). Note that since there are 3 input variables for this subsystem, there could be as many as 45 (3x5x3) 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* and then *peakness* is given more importance in setting up the fuzzy rules.

Table 4-4 provides the fuzzy rules for this system. Using this table, the system will decide the confidence in the bases called. For example, the first row of Table 4-4 indicates that: if confidence in *peakness* (C_p) is Low (L) and confidence in *height* (C_H) is Very Low (VL) and the confidence in *spacing* is Low (L), then the overall confidence in the base called (C_O) is Very Low (VL).

Table 4-4 If-then rules for *Overall Fuzzy Confidence System*

C_P	C_H	$C_{\Delta S}$	C_O
L	VL	L	VL
L	VL	M	VL
L	VL	H	VL
L	L	L	VL
L	L	M	VL
L	L	H	VL
L	M	L	L
L	M	M	L
L	M	H	L
L	H	L	M
L	H	M	M
L	H	H	M
L	VH	L	H
L	VH	M	H
L	VH	H	H
M	VL	L	VL
M	VL	M	VL

Table 4-4 continued.

M	VL	H	VL
M	L	L	VL
M	L	M	L
M	L	H	L
M	M	L	L
M	M	M	M
M	M	H	M
M	H	L	M
M	H	M	H
M	H	H	H
M	VH	L	H
M	VH	M	VH
M	VH	H	VH
H	VL	L	VL
H	VL	M	VL
H	VL	H	VL
H	L	L	L
H	L	M	L
H	L	H	M
H	M	L	M
H	M	M	M
H	M	H	M
H	H	L	H
H	H	M	H
H	H	H	H
H	VH	L	VH
H	VH	M	VH
H	VH	H	VH

All the implications, fuzzy operators within the antecedents, and implication aggregation follow the Mamdani model [17].

After rule evaluation, it is necessary to find the crisp output from the aggregate of all the results of the implication. We apply the center-of-gravity method because the aggregate implication results in a new fuzzy output set, while in fact we need a single crisp output. Applying the *maximum* function to all the resulting implications performs the aggregation.

5 Results and Analysis

This chapter illustrates the results from each fuzzy subsystem and the overall confidence values. It also describes the *TraceTools* software.

5.1 Confidence Values from each Fuzzy Subsystem

Results in this section are based on the raw data obtained from The ABI 3700 machine. To explain clearly, a part of the data, say six bases (ATCTCG) as shown in Figure 5-1 are described at each step. Note that the correctness of these bases was verified by correct contigs (ground truth).

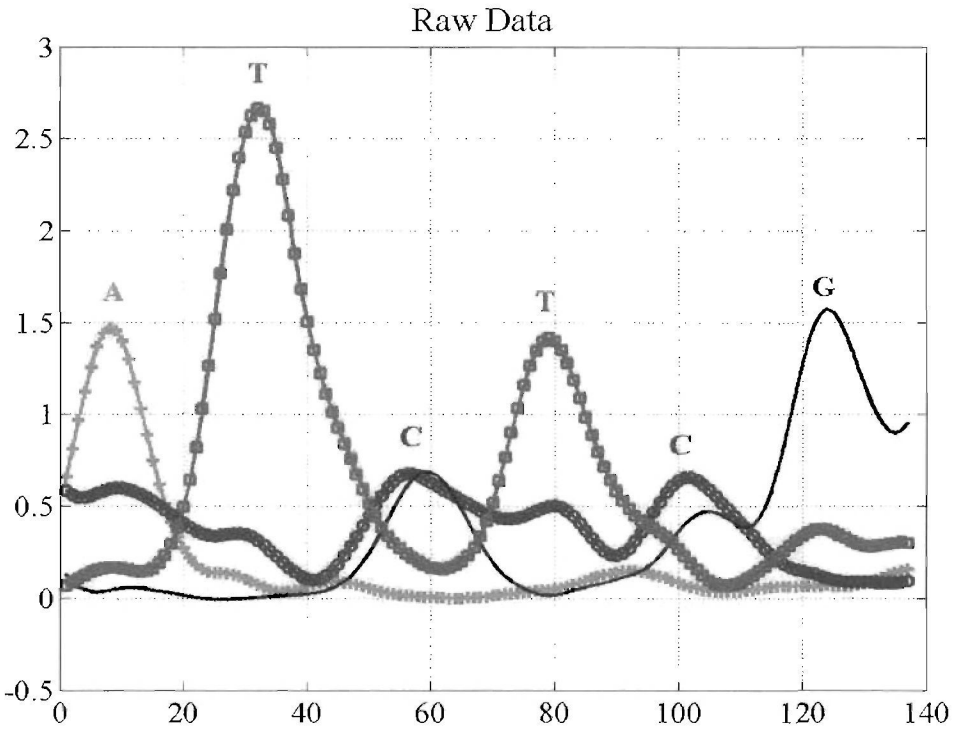

Figure 5-1 Raw data for 6 bases.

5.1.1 *Fuzzy Height* Subsystem

Figure 5-2 depicts the H_{called}, H_{2nd} as well as C_H for the six bases in the raw data in Figure 5-1. The actual values for each of the bases are shown in Table 5-1. For example, for the base T that is numbered 2, the normalized value of H_{called} is 0.991 while for H_{2nd} is 0.421. This distinction provides a high C_H confidence value of about 0.883 (solid line) for that base. Also, we can see for the next base C that is numbered 3, H_{called} is 0.644 and H_{2nd} is 0.604. There is not much height distinction between the first and the second candidates at that point. So the confidence value at that point will be less compared to the first base. Here we can see that the confidence value C_H is 0.124, which is very low. Similar analysis applies to other bases as well.

Table 5-1 Results table for *Fuzzy Height* Subsystem

Bases	H_{called}	H_{2nd}	C_H
A	0.889	0.560	**0.698**
T	0.991	0.421	**0.883**
C	0.644	0.604	**0.124**
T	0.954	0.606	**0.775**
C	0.696	0.531	**0.227**
G	0.952	0.485	**0.793**

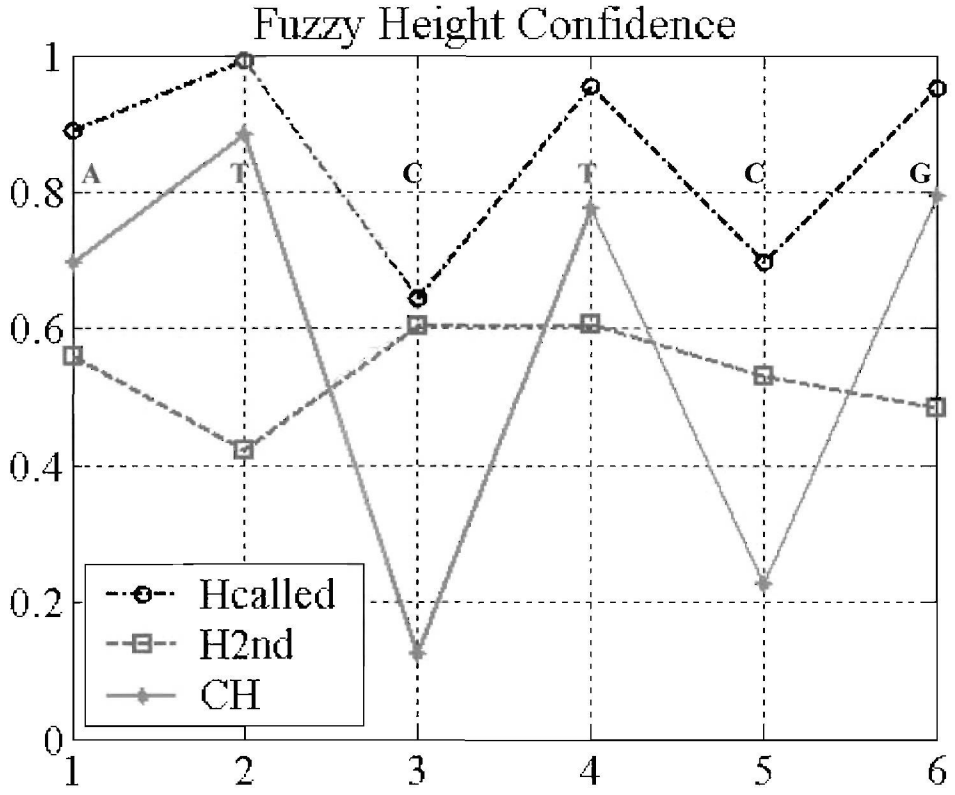

Figure 5-2 Results of *Fuzzy Height* Subsystem

5.1.2 *Fuzzy Peakness* Subsystem

Figure 5-3 depicts the P_{called}, P_{2nd} and C_P for the six bases in the raw data shown in Figure 5-1. Table 5-2 shows the corresponding values for the *Fuzzy Peakness* subsystem. Here, for the base T numbered 2, P_{called} is 0.999 while P_{2nd} is 0.478. This distinction provides a high confidence value C_P of 0.87 for that base. Also, for the next base C, the P_{called} is 0.794 while P_{2nd} is 0.838. It is very difficult to differentiate the two peaks at that point. So, we expect the confidence value based on the *peakness* to be low. The confidence value C_P is 0.548.

Table 5-2 Results table for *Fuzzy Peakness* Subsystem

Bases	P_{called}	P_{2nd}	C_P
A	0.998	0.361	**0.876**
T	0.999	0.478	**0.870**
C	0.794	0.838	**0.548**
T	0.999	0.721	**0.548**
C	0.930	0.665	**0.635**
G	0.999	0.618	**0.780**

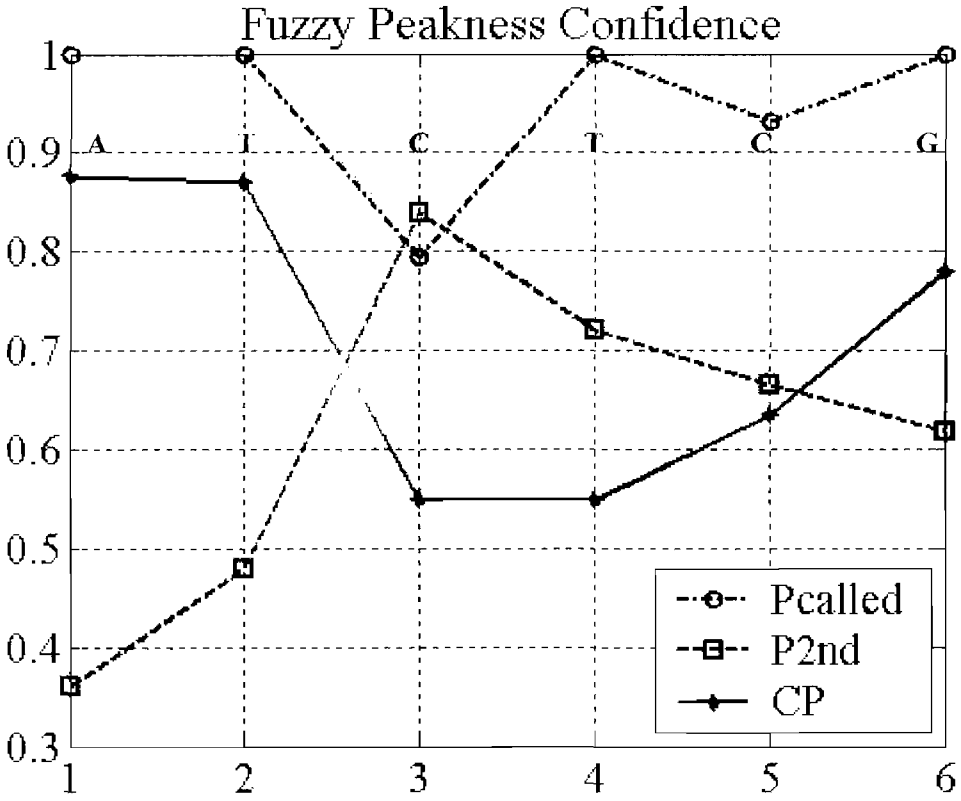

Figure 5-3 Results of *Fuzzy Peakness* Subsystem

5.1.3 *Fuzzy Spacing* Subsystem

Similarly, Figure 5-4 shows the ΔS_{next} and $\Delta S_{previous}$ for the six bases in the raw data and Table 5-3 shows the corresponding values. Here, for the same base T numbered 2, the values for both ΔS_{next} and $\Delta S_{previous}$ are 0.305. The actual spacing is so close to the predicted spacing that the confidence value $C_{\Delta S}$ will be high, 0.813, as seen in the figure. In fact, the confidence value for all bases, based on the spacing information alone, is high, in contrary to the other two measures, *height* and *peakness*.

Table 5-3 Results table for *Fuzzy Spacing* Subsystem

Bases	ΔS_{next}	$\Delta S_{previous}$	$C_{\Delta S}$
A	0.305	0.298	**0.824**
T	0.305	0.305	**0.813**
C	0.281	0.305	**0.825**
T	0.286	0.281	**0.825**
C	0.302	0.286	**0.825**
G	0.274	0.302	**0.825**

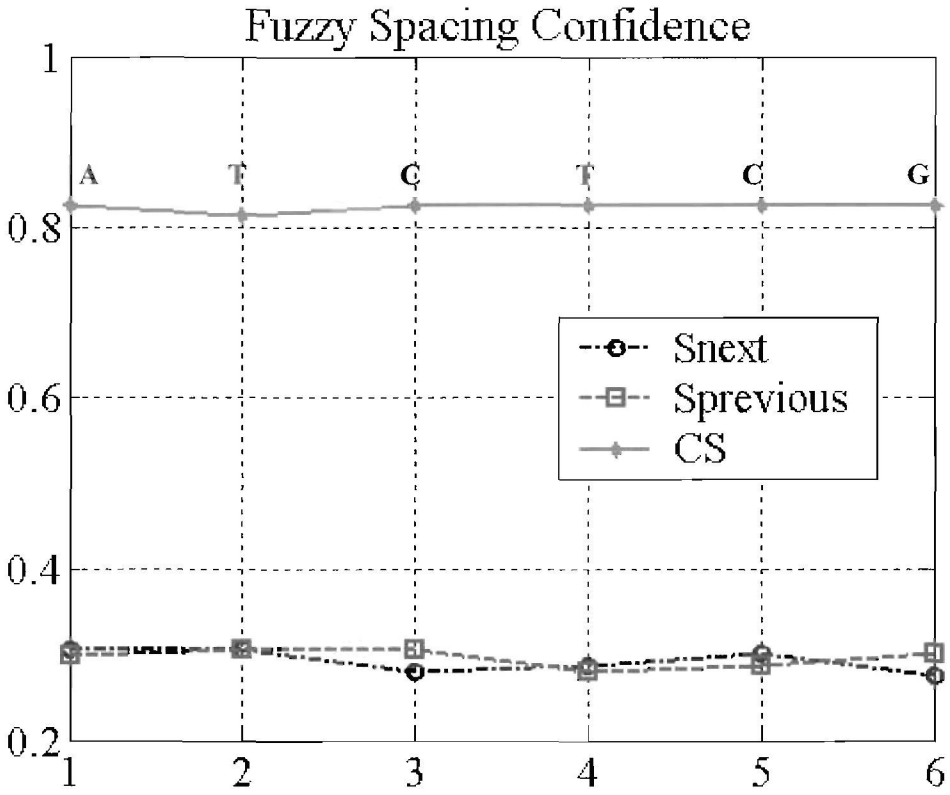

Figure 5-4 Results of *Fuzzy Spacing* Subsystem

5.2 *Overall Fuzzy Confidence* Subsystem

Combining the confidence values from the above three subsystems into the overall confidence value subsystem will provide the final confidence for the bases called. Figure 5-5 shows the overall confidence values (solid line) of the fuzzy system. The confidence values from each of the other subsystem are also shown in the graph. Table 5-4 shows the actual values corresponding to each system. In the first base A, when the confidence of each subsystem is high, we get a very high confidence value. For the base T that is numbered 2, where the confidence of each subsystem is high, a very high overall confidence value was obtained. For the base C numbered 3, which has a good confidence value for *spacing*, a very low confidence for *height,* and a medium confidence for *peakness*, an overall low confidence value was obtained. Note that the rules for the fuzzy system are designed in a way that more value is given to the *fuzzy height* confidence system hence, explaining why the overall confidence system and the height confidence system follow each other closely.

Table 5-4 Results table for *Fuzzy Overall Confidence*

Bases	C_P	C_H	C_{AS}	C_O
A	0.876	0.698	0.824	**0.837**
T	0.870	0.883	0.813	**0.921**
C	0.548	0.124	0.825	**0.124**
T	0.548	0.775	0.825	**0.775**
C	0.635	0.227	0.825	**0.219**
G	0.780	0.793	0.825	**0.901**

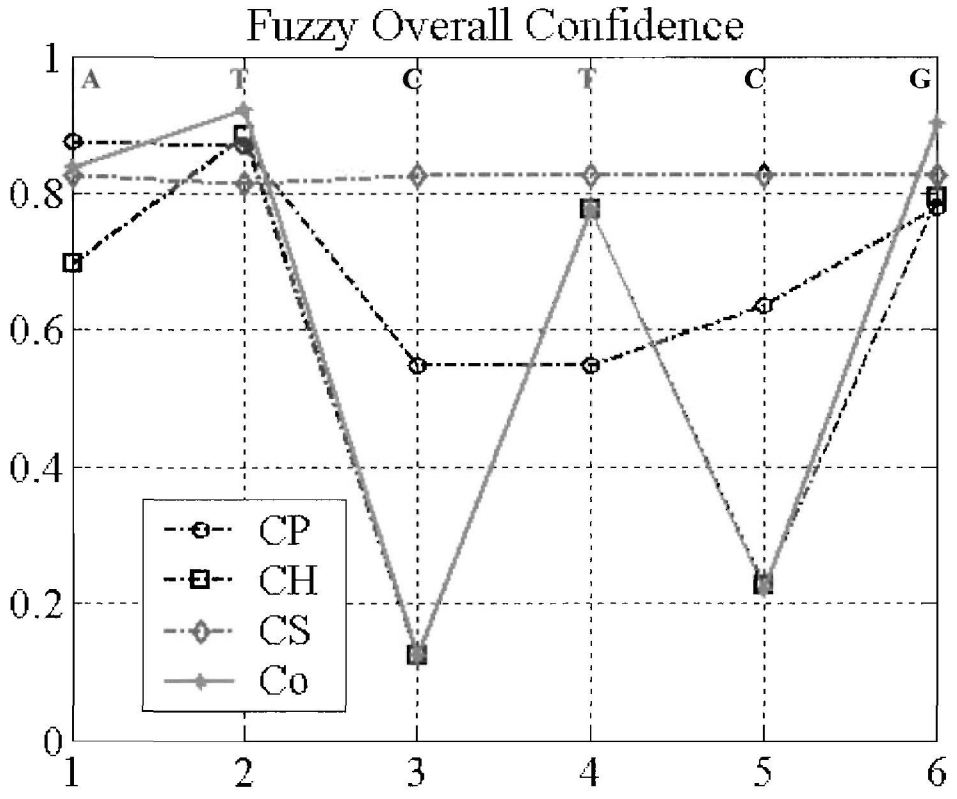

Figure 5-5 Results of *Overall Fuzzy Confidence* Subsystem

Considering the traces of Figure 5-1, and from a human operator point of view, the bases numbered 1,2 4 and 6 in the data can be called with higher confidences than the other two bases 3 and 5. The fuzzy model has indeed correctly assessed this observation. In fact, the fuzzy model has been tested on about 3000 files. Although there is no quantitative way of presenting the good "fit" of the model, visual inspection has indicated that the presented fuzzy confidence values follow the intuition of a human operator.

5.3 Confidence Values in *TraceTools* Software

Figure 5-6 shows a snap shot of *TraceTools*. This software is designed to process *ABI* 3700 chromatograms. *TraceTools* can display both the raw data (top window) and the processed data (bottom 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.

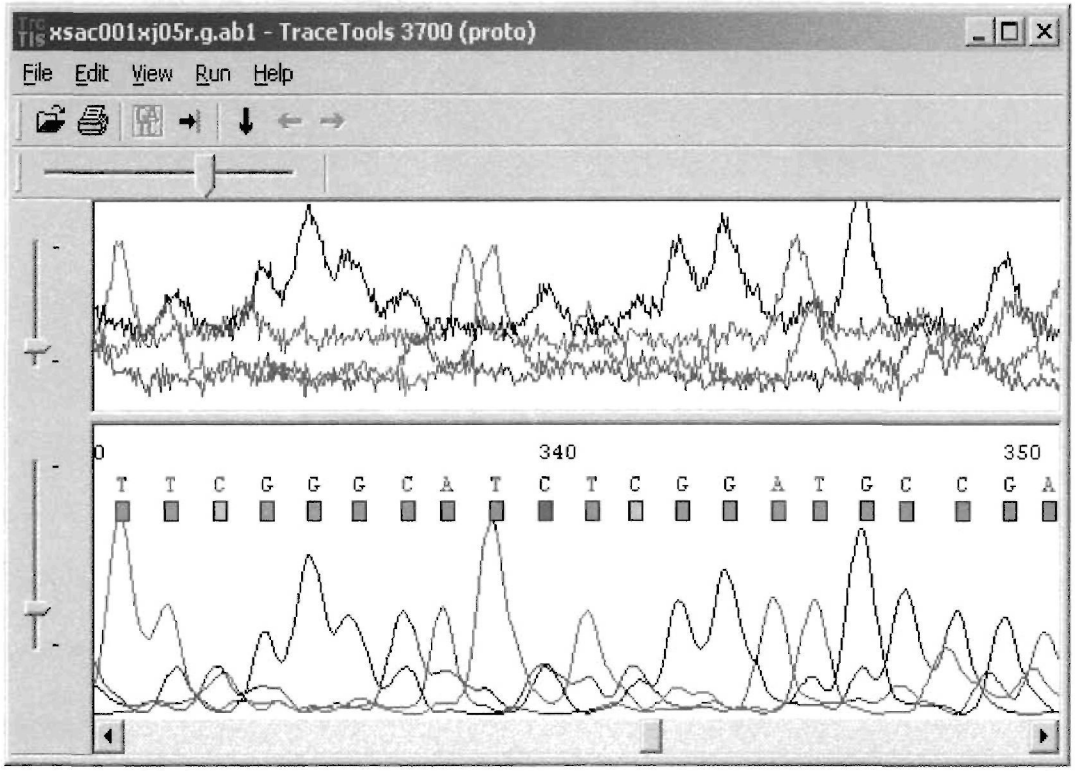

Figure 5-6 Snap shot of *Tracetools* software.

TraceTools displays the confidence measure associated with each base call through a color-coded rectangular bar. The color codes are as follows:

(Red): Very Low (yellow): Medium (green): High

The confidence values are indicated through rectangular bars. Green indicates highest confidence (50% or higher). The green box is further split into three parts to indicate confidence between 50 and 100%. If just the lowest part is colored green, the confidence value is between 50% and 60%.

If the bottom two parts are colored green, the confidence measure is between 60% and 75%. A fully colored green bar indicates a confidence measure between 75% and 100%. Yellow colored bar indicates 20% to 50% confidence. Red indicates not much confidence in the results obtained (20% or below), and recommends the user to manually make a base call.

The six bases (ATCTCG) around base 340 are the same bases that were discussed in the previous subsection. As seen from Figure 5-5, the two T bases have confidences of higher that 0.75; therefore full green bars present them. While the first and second C bases have confidence values 0.124 and 0.219, respectively. Thus, red and yellow bars indicate them, respectively. For easy evaluation purposes, the confidence values are multiplied by a factor of 10 before displaying in the software.

5.3.1 Results as shown in *Tracetools*

Figure 5-7 shows the display of the results in *TraceTools* obtained for the six bases considered in the previous sections. Here we can see that, a high confidence is shown in full green and a low confidence value in red and a medium confidence value in yellow color.

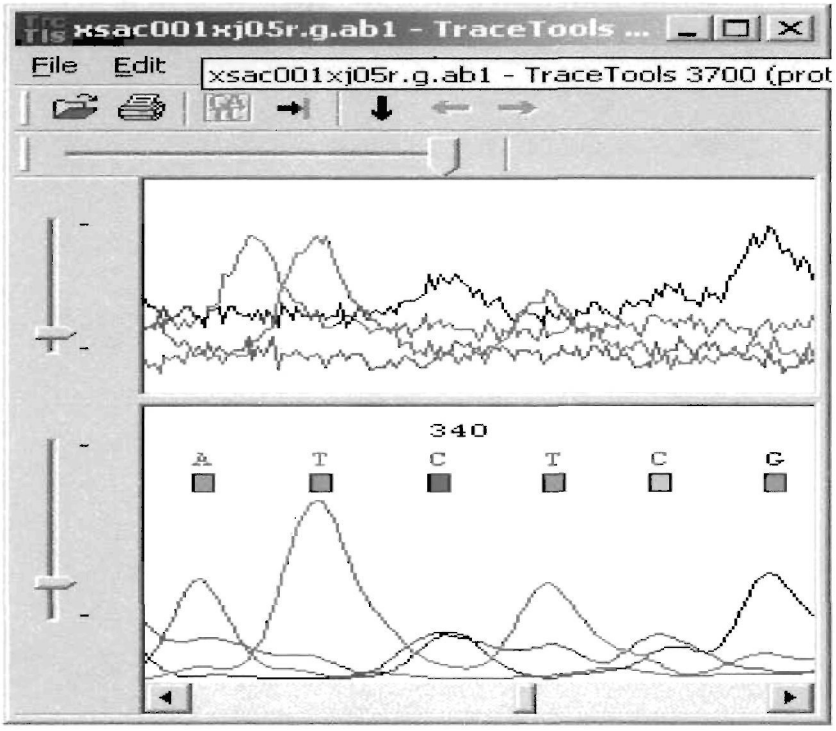

Figure 5-7 Results as shown in *Tracetools*

5.3.2 Comparison of Results with *Phred*

The only other technique that can be used for comparison of the results for this study is that of *Phred's or ABI's* quality value. Note that, although *Phred's* technique on quality values has been available for several years, ABI has just adapted the quality values in the ABI 3730 sequencing software. It is similar to the *Phred's* quality values. To show the performance of the confidence values of the proposed method with that of *Phred's*, the segment shown in Figure 5-7 is considered. Table 5-5 shows the quality values for *Phred* and the corresponding confidence values for *Tracetools.*

Table 5-5 Confidence values for *Phred* and *Tracetools*

	Confidence values for the bases called					
	A	T	C	T	C	G
Phred (Max value50)	20	12	14	13	15	22
Tracetools (Max value 10)	8.37	9.21	1.24	7.75	2.19	9.01

By looking at Figure 5-7, 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 9.21 (out of 10), 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 5-7 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.*

5.3.3 Data from 3730 DNA Analyzer

ABI 3730 is the successor to ABI 3700 DNA Analyzer. ABI predicts that this next-generation "production scale" machine will at least double the efficiency and quality of DNA sequencing data [18]. According to ABI, the advantages of the new machines fall into three categories; enhanced data quality, minimum reagent consumption, and fully automated production. The new machines feature sequence read lengths ranging from 550 bases to more than 1,000 nucleotides (using a 50-cm array) for the ABI 3730. *TraceTools* was able to read the new ABI 3730 data correctly, make the base calls and assign confidence values. Figure 5-8 shows the 3730 data in *TraceTools* and Figure 5-9 shows the same data as viewed by the new ABI software.

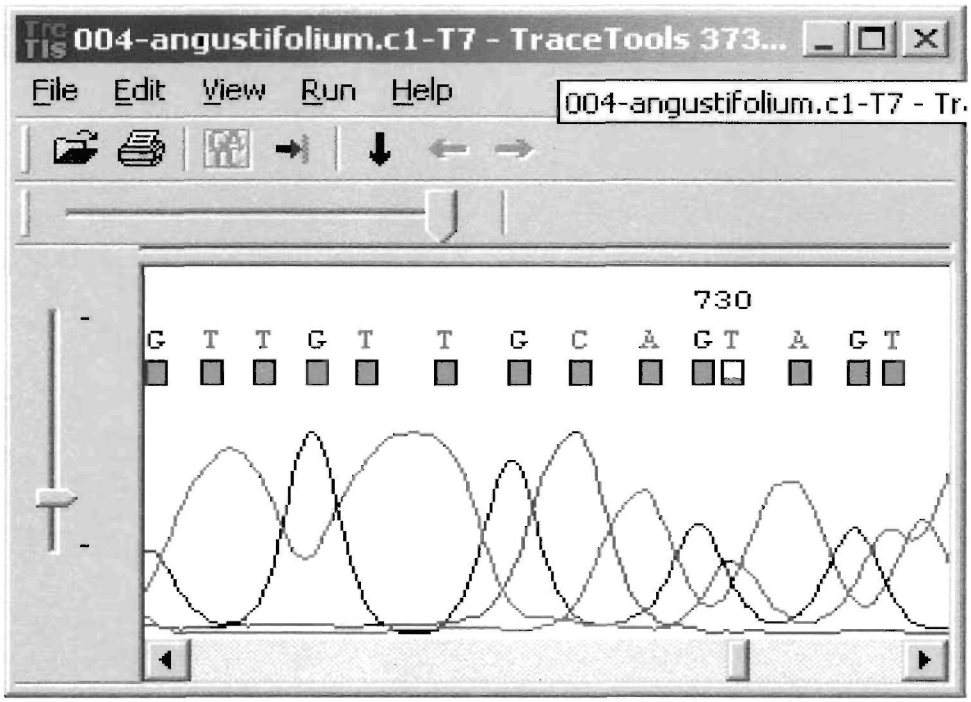

Figure 5-8 Results of 3730 data as shown in *Tracetools*

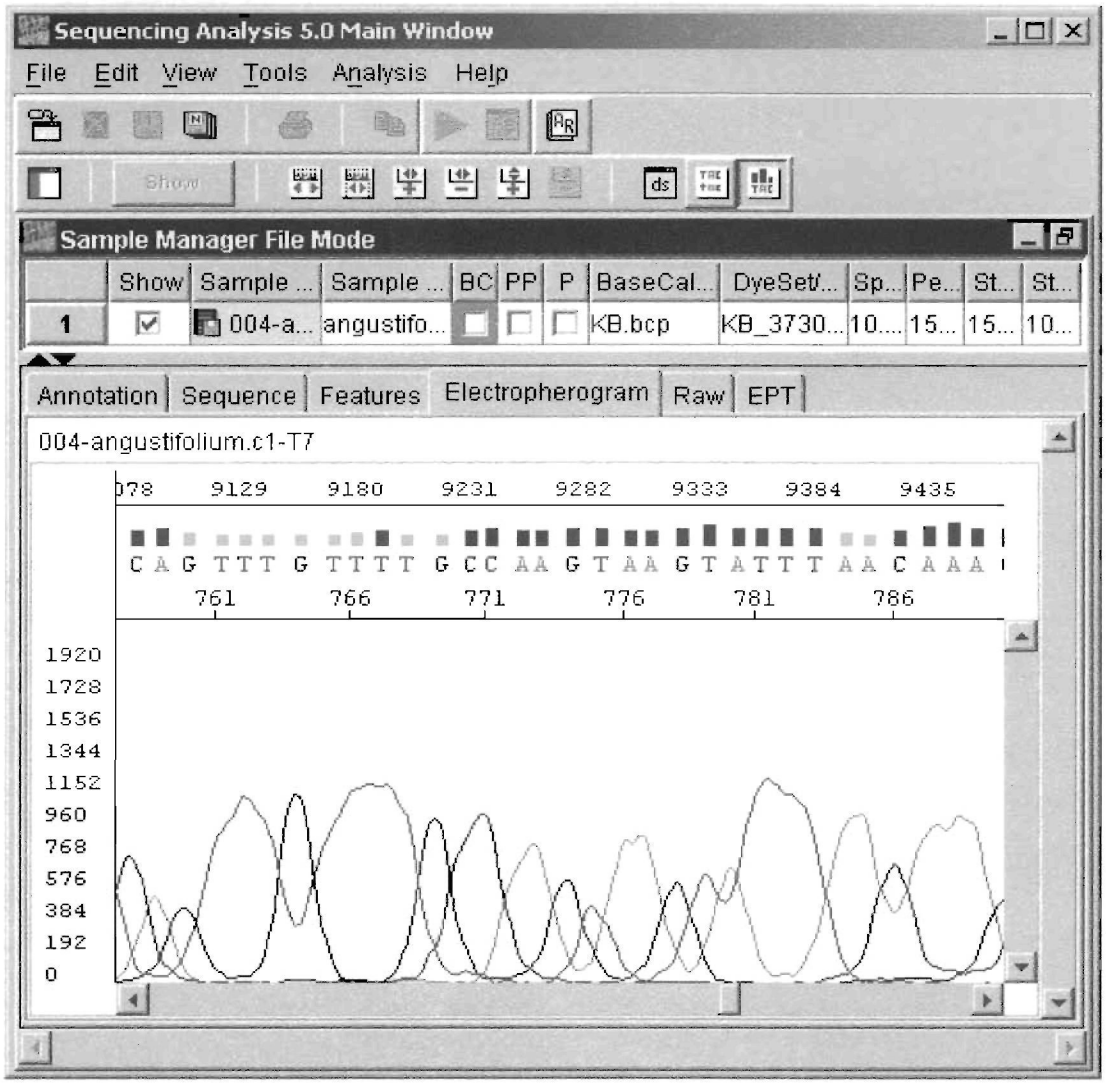

Figure 5-9 Results of 3730 as shown in ABI software

6 Confidence Values for Increasing the Accuracy

This chapter discusses on how the confidence system explained in this thesis can be used for improving the accuracy of the base caller.

6.1 Improvement in Accuracy

The confidence values generated using the fuzzy system can indeed be used as a measure for identifying the error areas in the DNA sequencing. By analyzing the areas of low confidence, one can provide solutions for improving the base calling on a local basis.

To test on the accuracy, four files are considered in which *Phred* has more accuracy than *TraceTools.* It is noticed that, in the areas where the fuzzy system was showing a low confidence, the base called by *TraceTools* was not correct. Each height and *peakness* at that point was considered then. It was noted that there was another winning candidate in that area. Considering this observation, other low confidence areas were looked into and second base calling was done based on the height, *peakness* and spacing in that region. Then an accuracy test was done on each of the four files considered. Table 6-1 describes the accuracy improvement in the four files considered.

Table 6-1 Improvement in accuracy based on confidence values

Flies	Accuracy before in %	Accuracy after in %
File 1	89.42	90.615
File 2	88.12	90.076
File 3	94.27	95.38
File 4	86.29	86.495

Based on these results, accuracy tests on all the 3000 ABI files were performed. We observed an accuracy improvement in the DNA base caller from 97.28% to 97.43%. Although this is not a high improvement in the accuracy, the increase should be noted. This explains the fact that the fuzzy confidence system developed in this thesis can be used for increasing the accuracy of a DNA base caller.

6.2 A Proposed Algorithm for Improving the Accuracy

Consider the Figure 6-1 that shows a sequence identified by the base caller *TraceTools*. The bases identified by the base caller in this area are AGAAAA. In the figure, we can easily see that there are 2 missed base calls and also an extra base. The extra base call is marked at data point 15259. Based on the contigs, the ground truth, the correct sequence in this region is AGGATAA. This is one of the cases where the base caller has made an error in identifying the correct bases.

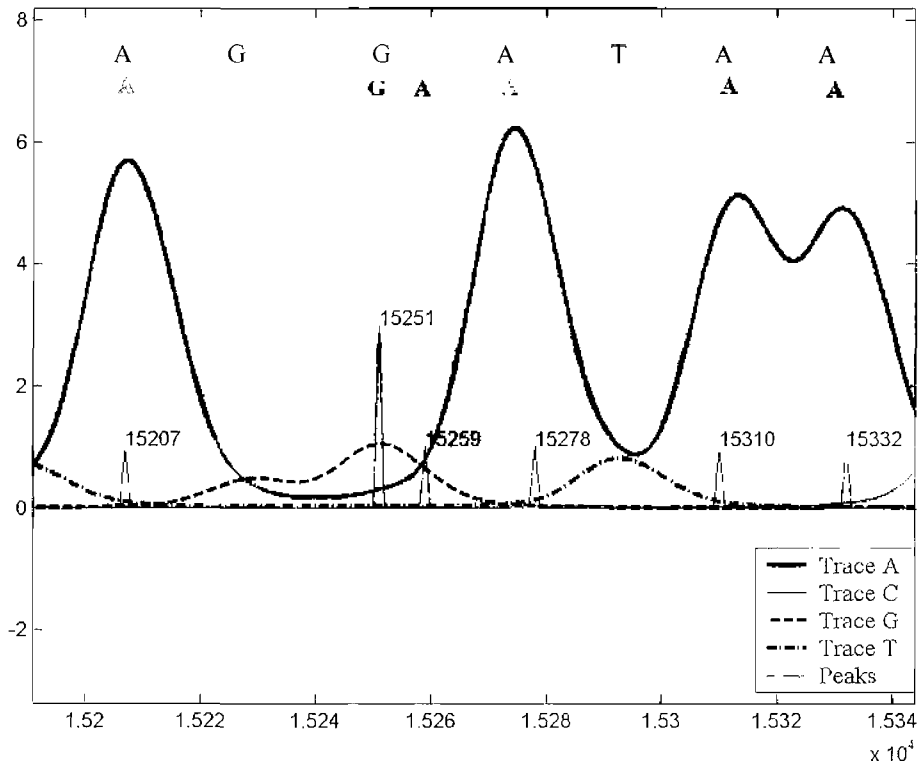

Figure 6-1. An example of a DNA sequence called by *TraceTools*

In *TraceTools*, the confidence value of the base 'A', called at data point 15259 has a very low confidence and the other bases have high confidence values. This shows that an algorithm could be introduced to identify the missing bases and discard the extra base. The algorithm in the next section can be followed.

6.2.1 Algorithm

Step 1: Identify the first low confidence base (base having a confidence value less than 20%) in the basecalling array.

Step 2: Consider two high confidence bases (bases with confidence value greater than 50%) to the either side of the base identified in step 1. Consider this as a window to perform the algorithm.

Step 3: From the raw data for *peakness*, identify the possible peaks in the window considered.

Step 4: Pass this array of peaks from step 3 into the confidence model to find the confidence values of each peak based on *peakness* and *height.* Rank the peaks based on the confidence values and eliminate peaks with low confidence. Consider the peak with the highest and second highest confidence.

Step 5: Calculate the predicted and actual spacing between the two bases considered. If the predicted spacing is less than the actual spacing then there is a possibility of finding a base between the two. Go to Step 6. If the predicted spacing is greater than the actual spacing, then there are no bases between them.

Step 6: If any peaks are found in between with good confidence value, then consider that as a peak candidate and repeat step 5 to the bases right and left of the peak considered.

Step 7: Check if the actual spacing is less than the average spacing for the file. If so, the peak considered is too close to the previous base called and so is not a base to be called. If not then go to step 5.

By this procedure, a missing base can be identified and also discard the extra base being called based on the spacing between the bases.

6.2.2 Calculation

(i) Using the confidence system, the highest confidence value was found for base A at 15274 and the second highest for base G at 15251.

The actual spacing between G and A = 23

The predicted spacing using the spacing model = 20.78

Since predicted < actual, possible peaks should be considered in between. One possible peak at 15257 was found, but was ignored due to low confidence value.

(ii) Consider from base G to the beginning of the window, i.e., A at 15207.

Actual spacing = 44

Predicted spacing = 31.826

Since predicted < actual, possible peaks should be considered. There were 3 possible candidates and only one had high confidence. This was a G at 15230. The spacing between the previous and the potential one is calculated.

Actual spacing = 23

Predicted spacing = 31.8

Since predicted > actual, no peaks or bases can exist in between these two bases. So the spacing between the potential to the next base is considered i.e., to the G at 15251.

Actual Spacing = 21

Predicted Spacing = 22.03.

Since predicted > actual, there is no bases in between. So the potential base G found at 15230 is now considered as a base.

(iii) Bases to the end of the window are also considered from base A at 15274 that was found in step 1. There is a base A at 15310.

Actual Spacing = 36

Predicted Spacing = 24.27

Since predicted < actual, possible peaks exist. Confidence values identify a base T at 15293. The spacing to the left and right of this base T is calculated and found that the predicted spacing > actual spacing and so no peaks exists in between.

At this point 4 bases are identified in between base A at 15207 and A at 15310. Those are: base G at 15230, base G at 15251, base A at 15274, and base T at 15293. Now the sequence becomes AGGATAA. This is same as the sequence identified by the contigs.

7 Conclusions and Future Work

7.1 Conclusions

The fuzzy confidence value system presented in this thesis is a powerful technique for providing the users of DNA sequencing software with a reliable measure of confidence in the bases called by the software. It will make the tedious correction and editing process much easier and faster. More importantly, since the results are reliable and true representation of the error areas, they can be used in the sequencing software as a reliability feedback measurement for further improvement. In other words, the confidence values can be used to automatically correct the base calling errors, hence, continually improving its performance.

7.2 Future Work

This thesis offers solutions to some of the challenges existing in DNA sequencing such as identifying the confidence values for the bases called. However, there are still many interesting issues in DNA sequencing that need future investigations.

Although the fuzzy system explained in this thesis can be used as a reliable representation of the areas, more fine-tuning can be done to make the fuzzy algorithm a proper tool for improving the accuracy of a DNA base caller. Fuzzy membership functions and the fuzzy rules can be investigated to

make this fuzzy system the best tool for DNA base callers. Genetic algorithms can be used to fine tune trapezoidal membership functions.

A different membership function like a Gaussian membership function can also be investigated. It is possible for the number of fuzzy sets to increase and the centers and widths of these Gaussian functions to change as well. This would take place in a tuning phase where one could identify numbers and centers of regions through clustering techniques such as fuzzy c-means clustering that can be performed on the input and output space. The results would directly relate to new member function locations and widths. These membership functions could be tuned further using neural networks, neuro-fuzzy, or genetic algorithms. In addition, if-then rules that we have established may be added or removed using neural-fuzzy techniques in an effort to further improve the model.

REFERENCES

[1] Bonfield, J.K. and Staden, R. (1995) "The application of numerical estimates of base calling accuracy to DNA sequencing projects." Nucleic Acids Research, 23, 1406-1410.

[2] Ewing, B. and Green, P. (1998) "Base-calling of automated sequencer traces using phred: II. Error probabilities." Genome Research, 8, 186-194.

[3] Luckey, J. A., Drossman, H., Kostichka, A.J., Mead, D. A., D'Cunha, J., Norris, T.B., and Smith, L. M. (August 1990) "High speed DNA sequencing by capillary electrophoresis", Nucleic Acids Research, 18, 15: 4417- 4421.

[4] Website: http://www2.carthage.edu/~pfaffle/hgp/ABI3700.html.

[5] Musavi, M.T., Domnisoru, C., Natarajan P., Varghese, R.S., Toothaker, M., Dawood, J., Ressom, H., Van Beneden, R., and Singer, P. (submitted 2004) "*TraceTools*: a new DNA base caller," submitted to Journal of DNA Sequencing and Mapping, January 2004.

[6] McNally, C., Domnisoru, C. and Musavi, M.T. (2002) "Building a DNA Database to Compare the Accuracy of Base Calling Programs." Proceedings of the METMBS International Conference on Mathematics and Engineering in Medicine and Biological Sciences, Las Vegas, Nevada, June 27, 2002, pp. 217-223.

[7] Domnisoru, C., Zhan, X., and Musavi, M.T. (2000) "Cross-talk Filtering in Four Dye Fluorescence-based DNA Sequencing." Electrophoresis. 21, 14: 2983-2989.

[8] Domnisoru, C., and Musavi, M. (2003) "Method for reducing cross talk within DNA data." United States of America Patent, No: 6,598,013.

[9] Website: http://mathworld.wolfram.com/RadiusofCurvature.html

[10] Finney, T. (2000) "Text Book: Thomas' Calculus," pp. 884-890.

[11] Tibbetts, C., and Bowling, J. (November 15, 1994) "Method and Apparatus for Automatic Nucleic Acid Sequence Determination," Vanderbilt University, Nashville, Tennessee, US Patent No: 5,365,455.

[12] Ewing, B., Hillier, L., Wendl, M. C., and Green P. (1998) "Base-calling of automated sequencer traces using Phred: I. Accuracy Assessment." Genome Research, 8, 175-185.

[13] Giddings, M., Rrumley, R., Haker, M., and Smith, L. (1993) "An adaptive, object oriented strategy for base calling in DNA sequence analysis." Nucleic Acids Research, Oxford University Press, 1993, pp. 4530-4540.

[14] Domnisoru, C., and Musavi, M.T. (2000) "Mechanical Shift and Base Spacing Modelling and Compensation for DNA Sequencing." Proceedings of the MS'2000 Modelling and Simulation, IASTED International Conference, Pittsburg, May 2000, pp. 470-476.

[15] French, B., Domnisoru, C., Ressom, H., and Musavi, M.T. (June 27, 2002) "Confidence Value Prediction of Called Genetic Bases Using a Fuzzy Prediction System." Proceedings of the International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS), Las Vegas, Nevada, USA, pp. 203-209.

[16] Yen, J. and Langari R. (1999) "Fuzzy Logic – Intelligence, Control and Information." Book: Published by Prentice Hall, Upper Saddle River, New Jersey, 1999.

[17] Mamdani, E.H. (1977) "Application of fuzzy logic to approximate reasoning using linguistic synthesis." IEEE Transactions on Computers, 26, 1182-1191.

[18] Website: http://www.bio-itworld.com/products/100902_abi.html.

BIOGRAPHY OF THE AUTHOR

Rency Susan Varghese was born in Mavelikkara, Kerala, India on January 3rd, 1977. She was raised in Trivandrum and graduated from Carmel Girls High School, Trivandrum in 1992 and obtained Pre-degree from Govt. College for Women, Trivandrum in 1994.

She entered the College of Engineering, Trivandrum (University of Kerala) and obtained her Bachelor's degree in Electrical and Electronics Engineering in 1999. After graduation she worked as Software-Engineer, at Satyam Computers Ltd, India for 1.5 years. Later in 2001 she joined Acclaim Systems, Inc as a Software Consultant.

In May 2002, she came to the United States of America and was enrolled for graduate study in Electrical Engineering at the University of Maine and served as Research Assistant in the Intelligent System Laboratory. Her current research interests include fuzzy logic and neural networks.

Rency is a member of IEEE and Sigma Xi Scientific Research Society. She is a candidate for the Master of Science degree in Electrical Engineering from The University of Maine in May, 2004.