A local processor for improving the accuracy of automated DNA sequencing

by

Murugathas Yuwaraj

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy,
Graduate Departments of Edward S. Rogers Sr. Department of Electrical and Computer Engineering and Institute of Biomaterials and Biomedical Engineering, University of Toronto

© Copyright Murugathas Yuwaraj 2001
The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author’s permission.

L’auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L’auteur conserve la propriété du droit d’auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-63645-3
A local processor for improving the accuracy of automated DNA sequencing

Doctor of Philosophy 2001

Murugathas Yuwaraj

Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto

Abstract

Electrophoresis-based automated identification of the Deoxyribo Nucleic Acid (DNA) sequence provides a means of detecting variations in the DNA of an organism at the nucleotide level. Accurate detection of these variations and the DNA sequence, in general, are limited by statistical variations in the DNA time-series and overlap of adjacent peaks due to limited resolution. In this thesis, a local maximum-likelihood processor (L-DNA), which is based on a statistical model of the DNA time-series is developed. The L-DNA processor generates a finite set of hypothesized sequences. For each hypothesized sequence, model parameters estimated from the observed waveform are used to evaluate the likelihood of the hypothesized sequence being consistent with the observed waveform. The hypothesis with the highest likelihood is selected as the local sequence in the observed region.

The performance of the L-DNA processor is evaluated using both simulated and real DNA data. Simulations show that in low resolution regions of the DNA data, which are observed towards the end of a sequencing run, the probability of error for the proposed processor is two orders of magnitude smaller than the commonly used Jansson-van Cittert iterative deconvolution method. The robustness of the processor is illustrated by sequentially extending the read length of a fragment of DNA sequence.

The ability of the processor to detect heterozygous sites and partial-mutation sites in low resolution regions is illustrated using 30 segments of DNA data where the proportion of the wild and mutant types is artificially controlled. The L-DNA
processor correctly identifies 80% of the data, while a commercial base-calling program correctly detects only 33% of the data. It should be noted that the commercial base-calling program is specifically designed for the sequencer that was used to acquire the above mentioned data.

The L-DNA processor provides a measure of confidence for each hypothesized sequence. This measure can be used to significantly reduce the amount of human intervention needed to edit results of automated DNA sequencing system in clinical applications.
Acknowledgments

Dr. M. Eizenman and Dr. S. Pasupathy, my supervisors, are thanked for their guidance and mentorship. Dr. Eizenman's interest in not only trying to steer the research in the right direction but also identify avenues for my personal career growth is immensely appreciated. I would like to express my sincere gratitude to Dr. R. Cobbold for introducing me to the field of Biomedical Engineering a decade ago as a summer student and providing valuable supervision ever since. I would also like to thank Dr. D. Hatzinakos for the expertise he provided as a committee member. Dr. F. Kschischang and Dr. S. Davies are thanked for providing a number of insights that helped me surpass many technical hurdles in my research work.

This work would not be complete without the large set of data which was generously made available by Visible Genetics Inc., Toronto. The staff at Visible Genetics, Chris Gabe, Jim Routliff and Rod Gilchrist are thanked for all the technical assistance, insights and most importantly, aiding in identifying key problems to attack in the automated DNA sequencing field.

At the Institute of Biomaterials and Biomedical Engineering, every student and staff is thanked for making the entire graduate student experience a memorable one. Ms. A. Mitchell, Prof. A.M. Dolan and Ms. S. Cherian are specially thanked for proving the infamous statement "it is possible!" when it comes to funding.

I would like to thank my parents for instilling in me the character to persevere and my wife for providing all the possible help and encouragement so that I could recognize this character in me. Last but not least, I would like to thank my baby boy for providing incentives to complete my graduate studies in a timely manner.
In fond memory of my mentor – a man who was an engineer at heart, a man who introduced me to the fascinating world of technology, a man who taught me to be proud of my roots – my grandfather, Mr. S. A. Rasanayagam.
Contents

Acknowledgments

List of Tables

List of Figures

List of Abbreviations and Symbols

1 Introduction

1.1 Motivation ........................................... 1
1.2 The DNA molecule and its role ....................... 2
1.3 Automated DNA sequencing .......................... 2
1.4 Automated DNA sequencing: challenges ............... 6
1.4.1 Resolution loss due to overlapping peaks .......... 6
1.4.2 Polymorphic sites ................................ 9
1.5 Resolving overlapping peaks: current algorithms .... 11
1.5.1 Deconvolution .................................. 12
1.5.2 Deconvolution techniques used in DNA sequencing .. 13
1.5.3 Other methods .................................. 17
1.6 Detecting polymorphic sites ......................... 20
1.6.1 Biological significance .......................... 20
## Contents

1.6.2 Current algorithms ................................................. 21  
1.7 Problem statement .................................................. 24  
1.8 Research approach .................................................. 24  
1.9 Dissertation organization .......................................... 25

2 Detection of DNA sequences in low-resolution data  ........................................................................... 27  
2.1 Introduction ............................................................... 27  
2.2 Analysis of van Cittert algorithm  .................................... 28  
2.3 Maximum likelihood detection of local sequence .......... 32  
2.3.1 Statistical Model ................................................... 33  
2.3.2 Local DNA Processor (L-DNA) .................................. 43  
2.3.3 Implementation ..................................................... 48

3 Performance Analysis .............................................................................................................................. 53  
3.1 Introduction ............................................................... 53  
3.2 Analysis using simulation ............................................. 54  
3.2.1 Implementation ..................................................... 54  
3.2.2 Results ................................................................. 56  
3.3 The L-DNA processor as a *proof reader* ....................... 57  
3.4 The L-DNA processor as a sequential decoder .............. 60  
3.5 Discussion ................................................................. 62

4 Heterozygote detection .............................................................................................................................. 64  
4.1 Introduction ............................................................... 64  
4.2 Adaptation of the L-DNA processor ............................... 65  
4.3 Computation of confidence measure for each hypothesis .... 67  
4.4 Use of L-DNA processor as a *proof reader* for detecting heterozygous sites ..................................... 68  
4.5 Performance of the L-DNA Processor ............................. 72  
4.5.1 Detection of heterozygous sites .................................. 73  
4.5.2 Detection of onset of mutations ................................. 79  
4.5.3 Generalized L-DNA processor ................................... 82
Contents

4.6 Discussion .............................................. 87

5 Summary and Future Research 89
  5.1 Thesis Summary ........................................ 89
  5.2 Thesis contributions .................................. 90
  5.3 Avenues for future research ......................... 91

Bibliography 93

A Implementation and analysis of JID algorithm 100
  A.1 Implementation details ............................... 100
  A.2 Empirical analysis of JID algorithm .............. 101

B Nomenclature used to represent DNA sequence 105
List of Tables

3.1 The L-DNA processor as a sequential decoder: sequence of most likely triplet of bases and the nearest competing hypotheses. .......................... 62

4.1 Top three hypothesized sequences identified by the L-DNA processor for the segments of data illustrated in Figures 4.2 and 4.3. Also illustrated in the table are the probabilities associated with each of the hypothesis being correct. .................................................. 72

4.2 Comparison of performance of the L-DNA processor with that of the commercial software for the 10 heterozygous cases. The top two hypotheses generated by the L-DNA processor for each data set and the relative probability of each of these hypothesis being correct is listed. The relative probabilities of the pairs of hypotheses (for each data set) do not add to unity since more than two hypotheses were considered for each data set. .......................................................... 75

4.3 Illustration of performance of L-DNA processor as a mutation detector for data with mixing ratio of 65:35. The subscripts of the hypothesized local sequences indicate the assumed mixing ratio. The performance of the commercial base-calling program is also presented. *In DataSet 13, the relative probability of TYT_{70:30} is 0.0640 and in DataSet 14, the relative probability of TYT_{70:30} is 0.0112. ........................................... 81
List of Tables

4.4 Illustration of performance of L-DNA processor as a mutation detector for data with mixing ratio of 75:25. The performance of the commercial base-calling program is also presented. 83

B.1 Nomenclature suggested by International Union of Biochemistry. 105
List of Figures

1.1 Illustration of a four lane automated DNA sequencing system. The sample sequence shown in this case is AGCTAT. The order of arrival of the bands, corresponding to each fragment, at the finishing line formed by a LASER beam determine the order of nucleotides in the sequence. Fluorescence is detected using photosensors (not shown) placed along the finishing line. .................................................. 4

1.2 Segment of time trace obtained from the Micro GeneBlaster$^{TM}$ sequencer. The chronological order of peaks provide the order in which bases appear in the DNA segment. In this case the sequence is AGCC-CCRCTTCATCKCAGT, where the heterozygous type R corresponds to the presence of A and G bases at the same position and K corresponds to the occurrence of T and G bases at the same position. Note the relative decrease in amplitude of heterozygous peaks. ........... 5
List of Figures

1.3 Illustration of some aspects of the DNA time series that limit accurate detection of nucleotides. Heterozygotes manifest as multiple small peaks at the same base position (see text for details). Top segment corresponds to fragments with 50 to 70 bases; middle segment corresponds to fragments with 140 to 160 bases and the bottom segment corresponds to 240 to 260 bases. The bottom segment exhibits broader peaks and hence lower resolution relative to the top segment with narrower peaks. .......................................................... 7

1.4 Comparison of segments of time series corresponding to heterozygous sites and sites that have undergone partial mutation. Member peaks are identified in the figure with arrows. The unequal heights in the case of partial-mutations result from the presence of wild and mutant types in uneven proportions. .......................................................... 8

1.5 A segment of four overlapping C peaks. Such an overlapping segment is referred to as a complex in this thesis. The local sequence is GAC-CCCACGTC. .......................................................... 10

1.6 Segment of DNA time trace illustrating a scenario where the two members (T and G peaks indicated with arrows) of a heterozygous site (K) are not perfectly aligned with respect to each other. The correct local sequence is TAKTACTG. Note that rest of the base-peaks are aligned reasonably well. .......................................................... 11

1.7 DNA time trace: a case where the strong G peak masks the presence of a smaller (heterozygous) G peak immediately adjacent to it. The position of the smaller G peak is indicated here with an arrow. The local sequence is CAGRCT, where R consists of G and A peaks. .... 12

1.8 Mapping of inputs and outputs of an ANN which serves as a signal conditioning processor (see Section 1.5.3 for description). Note that each connection (shown as lines) between processing nodes (shown as circles) has an associated weight value. The relative levels of the output nodes provide a measure of certainty associated with the respective base types. .......................................................... 26
List of Figures

2.1 Illustration of entire time series for A base type. Note that neither amplitude normalization nor constant offset removal has been done. 34
2.2 Segment of a DNA time-trace with large-scale trends removed. 35
2.3 Comparison of a proximal and a distal peak after the distal peak has been scaled horizontally. Both peaks have similar shape. 36
2.4 Comparison of an isolated DNA peak and the peak synthesized using (2.25). 37
2.5 Scatter plot of normalized amplitude estimates of peaks in the A lane of a typical sequencing run. 38
2.6 Peak amplitude covariance. 39
2.7 Scatter plot of estimated peak widths in the A lane of a typical sequencing run. Estimated widths of peaks from 25 base positions to 350 base positions are shown. 40
2.8 Illustration of presence of coloured noise. The T lane with sparse peaks (and hence prominent noise-only regions) is highlighted. Notice that the appearance of small noise peaks coincide with expected peak-times (indicated by the presence of peaks in other three lanes). 42
2.9 Covariance of noise peak amplitudes. 43
2.10 Peak-time jitter computed over a moving window of 40 bases. Percent jitter is defined as the ratio of standard deviation of peak-time and mean peak separation. 44
2.11 The L-DNA method. The particular hypothesized sequence with the highest likelihood is selected by the processor. 47
3.1 Comparison of DNA time-trace and synthesized time-trace. (a) DNA time-trace. The local sequence consists of a complex formed by the overlap of two G peaks. Hence the local sequence in this case is CGGA (b) time-trace synthesized using the statistics estimated from the DNA data. 55
3.2 Sensitivity of the processor to resolution of time series. 57
List of Figures

3.3 Sensitivity of the L-DNA processor to uncertainty in peak-time (jitter). $\sigma_t$ is the standard deviation of the peak-time described in (2.27). The estimated standard deviation of the probability of error is also shown. 58

3.4 Comparison of the JID and L-DNA methods using DNA time series. Top: Segment of the DNA time series with complex at approximately 2700. Bottom: The observed signal or the complex (solid line), deconvolved signal generated by JID method (dash dot) and the three estimated components obtained from L-DNA method, corresponding to the hypothesis CCC (dashed lines). 59

3.5 Illustration of the L-DNA processor as a sequential decoder of a segment of data towards the end of a run where significant overlapping of peaks occur. 61

4.1 Steps involved in detecting the local sequence in a complex using the L-DNA method. The particular hypothesized sequence with the highest likelihood (or the lowest cost) is selected by the processor. It should be noted that the set proportion of heterozygous peaks is hypothesis dependent. 66

4.2 A segment of data with a heterozygous base type which is difficult to identify due to the presence of overlapping homozygous base of the same type. In this case the correct local sequence is GRC. The R base type is formed by the presence of 50% of G nucleotides and 50% of A nucleotides in the population DNA that was sequenced. 69

4.3 Second example used to illustrate the use of L-DNA processor to identify heterozygous sites. In this case the correct local sequence is ARC. The heavy tail of the complex implies that the peaks in this lane (A lane) are highly asymmetric. 70
List of Figures

4.4 Results of processing data in Figure 4.2 using the L-DNA processor. The solid line indicates the observed G trace and the dashed lines are the two estimated components in the G lane that correspond to the most likely hypothesized sequence. The result obtained using a commonly used iterative deconvolution method (Janssen-van Cittert iterative deconvolution (JID) method) is also shown. .......................... 73

4.5 Results of processing data in Figure 4.3 using the L-DNA processor. The solid line indicates the observed complex in the A lane and the dashed lines are the two estimated components in the A lane that correspond to the most likely hypothesized sequence. Also shown is the result obtained using the Janssen-van Cittert iterative deconvolution algorithm (JID). ........................................ 74

4.6 Results of processing data in Figure 4.3 using the L-DNA processor assuming that the peak shape is symmetric. Notice that the resulting estimates are almost equal in amplitude though the first component is supposed to be a homozygous type and the second component is supposed to be a heterozygous type. ........................................ 76

4.7 Illustration of two known polymorphic sites used to assess the performance of the L-DNA processor. The complexes are located at approximate base position of 250 and the complete read length is approximately 300. (a) Here the heterozygous type is K. For a window of 3 nucleotides that include the complex, the correct local sequence is KTA. (b) Here the heterozygous type is Y. For a window of 3 nucleotides that include the complex, the correct local sequence is TYT. .......................... 77

4.8 Detection of constituent peaks in the complexes illustrated in Figure 4.7. (a) Here the observed T lane and the estimated components associated with the detected sequence in the complex shown in Figure 4.7(a) are illustrated. (b) Here the observed T lane and the estimated components that constitute the complex are shown for the segment illustrated in Figure 4.7(b). ........................................ 78
List of Figures

4.9 Performance summary of the L-DNA processor. Pairs of base types in parenthesis indicate the two types of bases that together constitute a heterozygous type present at a given base position. The ratios in parenthesis refer to the proportions in which the two nucleotides were mixed to generate the heterozygous type. ........................................ 84

4.10 Generalized L-DNA processor: Implementation of a combined structure to detect not only possible polymorphic sites but also classify them as either pure heterozygotes or sites exhibiting early stages of mutation in various proportions. ........................................ 86

A.1 Illustration of peak sharpening property of JID algorithm. ....... 102

A.2 Comparison of Jansson-van Cittert algorithm and the general van Cittert algorithm. ........................................ 103

A.3 Performance of JID algorithm with respect to resolution of signals. Here, resolution R is defined as the ratio of peak width and peak separation. ........................................ 104
List of Abbreviations & Symbols

### ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>adenine</td>
</tr>
<tr>
<td>ANN</td>
<td>artificial neural network</td>
</tr>
<tr>
<td>C</td>
<td>cytosine</td>
</tr>
<tr>
<td>dATP</td>
<td>deoxyadenosine triphosphate</td>
</tr>
<tr>
<td>dCTP</td>
<td>deoxycytidine triphosphate</td>
</tr>
<tr>
<td>dGTP</td>
<td>deoxyguanosine triphosphate</td>
</tr>
<tr>
<td>dTTP</td>
<td>deoxythymidine triphosphate</td>
</tr>
<tr>
<td>dNTP</td>
<td>deoxynucleotide triphosphate</td>
</tr>
<tr>
<td>ddATP</td>
<td>dideoxyadenosine triphosphate</td>
</tr>
<tr>
<td>ddCTP</td>
<td>dideoxycytidine triphosphate</td>
</tr>
<tr>
<td>ddGTP</td>
<td>dideoxyguanosine triphosphate</td>
</tr>
<tr>
<td>ddTTP</td>
<td>dideoxythymidine triphosphate</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>FWHM</td>
<td>full width at half maximum</td>
</tr>
<tr>
<td>G</td>
<td>guanine</td>
</tr>
<tr>
<td>HLA</td>
<td>human leukocyte-associated antigens</td>
</tr>
<tr>
<td>ISI</td>
<td>inter-symbol interference</td>
</tr>
<tr>
<td>JID</td>
<td>Jansson-van Cittert iterative deconvolution</td>
</tr>
<tr>
<td>L-DNA</td>
<td>local DNA processor</td>
</tr>
<tr>
<td>ML</td>
<td>maximum likelihood</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>pdf</td>
<td>probability density function</td>
</tr>
<tr>
<td>PSF</td>
<td>point spread function</td>
</tr>
<tr>
<td>ssDNA</td>
<td>single stranded DNA</td>
</tr>
<tr>
<td>SNR</td>
<td>signal to noise ratio</td>
</tr>
<tr>
<td>T</td>
<td>tyrosine</td>
</tr>
</tbody>
</table>
List of Abbreviations & Symbols

SELECTED SYMBOLS

\( y \) observation vector
\( x \) original signal in deconvolution methods
\( C \) constant of proportionality relating mobility and molecular length
\( D \) diffusion coefficient
\( \hat{D} \) distortion operator
\( n \) noise
\( g \) point spread function in deconvolution and peak shape in statistical model
\( \beta \) relaxation factor in iterative deconvolution
\( C \) constraint operator
\( \gamma \) maximum allowable amplitude in JID algorithm
\( x^{(k)} \) estimated original signal at k-th iteration
\( X \) Fourier transform of original signal vector
\( Y \) Fourier transform of observation vector
\( G \) Fourier transform of point spread function
\( N \) Fourier transform of the noise vector
\( \lambda \) eigenvalue
\( t_i \) peak time for base position i
\( \text{pw} \) peak width
\( g_i() \) generic pulse shape
\( g_{t_i,t_i} \) pulse shape peaking at \( t_i \) evaluated at \( k \)
\( \delta_{i,j} \) Kronecker delta function
\( \sum \) summation
\( \mu_T \) mean peak separation
\( \phi \) jitter
\( S \) local DNA sequence
\( p(y|x) \) conditional pdf
\( l \) vector of peak amplitudes and noise amplitudes
\( a \) vector of peak amplitude
\( J_i \) percent jitter
\( R \) resolution ratio
CHAPTER 1

Introduction

In the post-human genome era, the challenge lies in the accurate detection and subsequent interpretation of subtle variations that exist throughout the human genome. These variations can mean the difference between susceptibility and resistance to a particular disease, compatibility and rejection for a particular tissue transplant, or presence and absence of an onset of pathogen attack. This chapter introduces the concept of DNA variation, its biological significance and factors that limit their accurate detection. Benefits and drawbacks of current detection methods are also identified. The research approach adopted in this work is presented in light of the insight gained from previous work and exploration of the data. Finally, an overview of this thesis is provided.

1.1 Motivation

The successful completion of the human genome project has given birth to a whole new area which involves the interpretation of genes and their clinical implications. The clinical acceptance of DNA tests would, in turn, demand accurate and efficient
interpretation of each patient’s DNA sequence. Currently, substantial human intervention is required to edit errors in sequences read by automated DNA sequencers. Hence, there is a clear need from a clinical perspective to automate the proof reading task. If a statistical frame work could be adopted to automate this task, then a significant benefit would be the ability to provide relative probabilities associated with each of the decisions of the automated proof reader. Further, significant time, and hence cost, savings would result from the elimination of the manual proof reading phase. Armed with quantitative confidence measures, the human expert will be at a better position to draw higher level conclusions from the read DNA sequence.

1.2 The DNA molecule and its role

The Deoxyribonucleic acid (DNA) molecule is a double helix composed of nucleotides arranged in a particular sequence that is unique to each individual. Genetic information is encoded according to the particular nitrogenous base (Adenine (A), Cytosine (C), Guanine (G), or Thymine (T)) present in each nucleotide. Each consecutive triplet of bases either codes for one of twenty amino acids or it codes for control information. The amino acids and the control information are used in the synthesis of structural and enzymatic proteins of the cell. The sequence of nucleic acids that carries the information representing a particular protein is called a gene. The enzymes, in turn, control the formation of other structures in the cells and determine the functional activity of the cell by regulating various metabolic reactions. Through this regulation at the cellular level, the DNA determines the structure and function of the body as a whole [1].

1.3 Automated DNA sequencing

In the Sanger method [2] for DNA sequencing, a polymerase enzyme is used to prepare a set of partial copies of the original single stranded DNA (ssDNA) molecule. The use of a primer molecule ensures that all subsequences start at the same location. In addition to the dATP, dCTP, dGTP and dTTP substrates needed to make mul-
multiple copies, additional terminator molecules (ddATP, ddCTP, ddGTP or ddTTP) are included in the synthesis process. The terminator molecules compete with the corresponding normal substrate molecules for incorporation into the growing copies. When the terminator molecules are incorporated into the sequence, the copying process is halted and the length of the partial copy is fixed. For example, reaction carried out in the presence of ddATP results in the generation of all possible subsequences that end in an A nucleotide. Since incorporation of the terminator molecule occurs by chance, the final result contains all possible subsequences of the original ssDNA. Thus, the lengths (or molecular weight) of the partial copies correspond to the position of a particular base type. Electrophoresis is then used to separate these negatively charged DNA molecules. The samples are placed at the top end of a gel and a voltage is applied across the gel. Since the small DNA molecules (shorter fragments) migrate faster down the gel than the larger molecules (longer fragments), separation is achieved based on size. Since molecules of similar length migrate together, they appear as migrating bands in the electrophoresis gel. In automated DNA sequencing, fluorescent molecules are attached to either the terminator molecule (dye-terminator chemistry) or the primer molecule (dye-primer chemistry) of each copy and they are detected using a photo sensor as the molecules migrate past a finishing line formed by a laser beam. Such a system is illustrated in Figure 1.1. The order in which signals appear at each of the sensors determine the order of the bases in the sequence under consideration (see Figure 1.2 for a typical time trace obtained using an automated DNA sequencer). The time traces corresponding to each of the four base types will be referred to as lanes in this thesis. For example, the A time trace will be referred to as the A lane. The terms, time-series and trace, will be also used interchangeably.

Two sites, known as heterozygous sites, are also illustrated in Figure 1.2. In heterozygous sites, one of two nucleotides are present at the same base position in each DNA fragment. Hence, the DNA time trace exhibits two peaks at the same position. A heterozygous site arises due to the contribution of different nucleotides for the same base position from the two parents. This leads to half of the population of samples containing one base-type and the other half of the population containing another
Figure 1.1: Illustration of a four lane automated DNA sequencing system. The sample sequence shown in this case is AGCTAT. The order of arrival of the bands, corresponding to each fragment, at the finishing line formed by a LASER beam determine the order of nucleotides in the sequence. Fluorescence is detected using photosensors (not shown) placed along the finishing line.
Figure 1.2: Segment of time trace obtained from the Micro GeneBlaster™ sequencer. The chronological order of peaks provide the order in which bases appear in the DNA segment. In this case the sequence is AGCCCCRCTTCATCKCAGT, where the heterozygous type R corresponds to the presence of A and G bases at the same position and K corresponds to the occurrence of T and G bases at the same position. Note the relative decrease in amplitude of heterozygous peaks.
base-type. Since the total population size of each fragment of DNA is approximately the same, the two peaks appear as approximately half the height of homozygous sites (where only one base type is present in the entire population). The significance of heterozygous sites are explained in Section 1.6.1.

1.4 Automated DNA sequencing: challenges

Automated detection of the sequence of nucleotides from the time trace obtained by DNA sequencing is limited by such factors as variability in amplitudes of the peaks, variability in the peak-times, spurious peaks occurring due to artifacts in the reactions and loss of resolution. Figure 1.3 illustrates three segments of the DNA time series that were acquired in the beginning, middle and end of the electrophoresis process. Though the peaks appear well separated in the initial segment (top), gradual degradation in resolution due to broadening of peaks is evident with the progression of time (or increase in length of the DNA segment). Among these factors, the errors associated with loss of resolution have been attributed to more than half of the total errors in DNA base-calling [3, 4, 5]. Another aspect that is particularly challenging is the identification of sites that are heterozygous in nature (described previously) and sites that have undergone partial mutation. In the latter case, the time-trace resembles that of a heterozygous type but the heights of the wild and the mutant nucleotide peaks may not be equal. The relative proportion of the heights are determined by the extent of the mutation. Figure 1.4 illustrates the time series corresponding to these two similar cases. In this thesis, sites in a DNA sequence which exhibit the presence of more than one type of nucleotide, such as heterozygous sites and sites with partial-mutation, are referred as polymorphic sites. These terms will be used interchangeably.

1.4.1 Resolution loss due to overlapping peaks

The loss in resolution result from broadening of peaks while the mean peak separation remains almost constant. This, in turn, leads to overlapping. The peak width is determined by the initial peak width on loading of the sample into the electrophoresis
Figure 1.3: Illustration of some aspects of the DNA time series that limit accurate detection of nucleotides. Heterozygotes manifest as multiple small peaks at the same base position (see text for details). Top segment corresponds to fragments with 50 to 70 bases; middle segment corresponds to fragments with 140 to 160 bases and the bottom segment corresponds to 240 to 260 bases. The bottom segment exhibits broader peaks and hence lower resolution relative to the top segment with narrower peaks.
Figure 1.4: Comparison of segments of time series corresponding to heterozygous sites and sites that have undergone partial mutation. Member peaks are identified in the figure with arrows. The unequal heights in the case of partial-mutations result from the presence of wild and mutant types in uneven proportions.

gel and the additional increments due to diffusion and dispersion since loading \([6]\). The diffusion refers to the effect present in the absence of an electric field and dispersion refers to the time-dependent component present when an electric field exists. The time-dependent spatial width, \(\Delta x\), of a band\(^1\) in the electrophoresis process can be modeled as \([7]\):

\[
\Delta x_D^2(t) = \Delta x_0^2 + 2Dt
\]  \hspace{1cm} (1.1)

where \(t\) is time since loading of samples and application of an electric field, \(\Delta x_0\) is the width of the band at the time of loading and \(D\) is the diffusion coefficient that collectively represents diffusion and dispersion. The width of a peak \((w)\) at a particular time is then:

\[
w = \Delta x_D / v
\]  \hspace{1cm} (1.2)

\(^1\)Molecules of similar length which migrate together in an electrophoresis gel appear as a band.
where \( v \) is the speed of the peak (band) in the gel. The biased reptation model [7] for migration of DNA molecules in a gel suggests that the velocity of the peak in the gel (mobility) is inversely proportional to the molecular length and hence base number, which is described by \( i \). Hence,

\[
v = \frac{C}{i}
\]  \hspace{1cm} (1.3)

where \( C \) is a constant of proportionality. If \( L \) is the length of the gel, a peak passes the sensor at the bottom of the gel at \( t = \frac{L}{v} = \frac{Li}{C} \). Hence, the width of the \( i \)-th peak at the sensor can be rewritten as:

\[
w(i) = \frac{i}{C} \sqrt{\Delta x_0^2 + \frac{2DLi}{C}}
\]  \hspace{1cm} (1.4)

This expression implies that the width of a peak grows at a rate that is between \( i \) and \( i^{3/2} \). It is clear that broadening of peaks, and hence loss of resolution, are more prevalent as the base number increases. Such overlapping peaks are shown in Figure 1.5. The broad peak resulting from overlapping of peaks is referred in this thesis as a complex.

1.4.2 Polymorphic sites

Though DNA sequence analysis provides a means of automating the detection of DNA sequence, polymorphic sites are particularly hard to accurately identify for the following reasons: (a) larger variation in the amplitude of heterozygous peaks relative to homozygous peaks, (b) temporary loss of alignment between the four lanes near polymorphic sites, and (c) loss in resolution due to bias introduced by adjacent strong peaks (homozygous peaks) of the same base type.

Variability in fluorescence signals arise due to nonuniform incorporation of fluorescent markers. Such variability is reduced to certain extent when the fluorescent markers are incorporated in the primer molecules instead of the terminator molecules. Yet, the random nature of chemical reactions results in variation in amplitude levels of the measured signals. Such variations in signal levels in the same lane are clearly
Figure 1.5: A segment of four overlapping C peaks. Such an overlapping segment is referred to as a complex in this thesis. The local sequence is GACCCCACGTC.

visible in the time trace illustrated in Figure 1.2.

In polymorphic sites, the terminus base types of the two members differ, even though the lengths of the two strands are the same. Hence, the mobility of the two types of molecules may differ. This difference in mobility causes the two peaks corresponding to the two members of the heterozygous to arrive at the finishing line at slightly different instances. In other words, the two peaks are imperfectly aligned with respect to each other. This misalignment is illustrated in Figure 1.6.

The loss in resolution, described in Section 1.4.1, further complicates the detection of polymorphic sites. The presence of a homozygous peak, particularly towards the end of a sequencing run, immediately before or after a heterozygous member from the same lane often obscures or biases the smaller heterozygous peak. In Figure 1.7, the presence of a full-height G peak immediately before a half-height G completely obscures the latter peak.
Sections 1.5 and 1.6 provide an overview of past attempts at circumventing limitations in accurate determination of DNA sequences due to overlapping peaks and polymorphic sites.

1.5 Resolving overlapping peaks: current algorithms

Efforts to resolve overlapping peaks have been primarily focused on the use of deconvolution algorithms. These methods are described in Section 1.5.2. Some attempts were also made to use Artificial Neural Networks, peak shape fitting methods and statistical model based methods. These approaches are analyzed in Section 1.5.3. First, a general introduction to the deconvolution problem is given.

Figure 1.6: Segment of DNA time trace illustrating a scenario where the two members (T and G peaks indicated with arrows) of a heterozygous site (K) are not perfectly aligned with respect to each other. The correct local sequence is TAKTACTG. Note that rest of the base-peaks are aligned reasonably well.
Figure 1.7: DNA time trace: a case where the strong G peak masks the presence of a smaller (heterozygous) G peak immediately adjacent to it. The position of the smaller G peak is indicated here with an arrow. The local sequence is CAGRCT, where R consists of G and A peaks.

1.5.1 Deconvolution

One approach to restoring the underlying peak shapes from overlapping signals is to treat the observed signal as a convolution of an ideal signal with some smearing function which serves to broaden and overlap peaks. The mathematical representation of this process is:

$$y = Dx$$

(1.5)

where $x$ is the unknown input signal, $y$ is the observed or measured signal and $D$ is a distortion operator. Signal restoration involves the recovery of $x$ given $y$ and $D$. The straightforward approach to solving this problem is to find the inverse operator $D^{-1}$.
of $D$ such that

$$x = D^{-1}y$$  \hfill (1.6)

There are many limitations to direct implementation of inverse operators for signal restoration. If $y$ is known imprecisely, for example due to additive noise, then result of applying the inverse operator may differ significantly from the true solution. Also, if the distortion operator $D$ is such that many inputs $x$ are mapped to the same output $y$, an inverse operator does not exist. If $D$ is a bandlimiting operator (such as a low pass filter), then it results in many-to-one mapping and the inverse does not exist. In such cases, prior knowledge of the properties of $x$ will aid in eliminating some of the ambiguities.

For reasons cited above, several alternative schemes have been developed which attempt to estimate an approximation to $x$ without necessarily inverting the distortion operator. Since a plethora of techniques exist for deconvolution (see [8, 9] for a good sampling of methodologies) only those techniques that have been previously explored for DNA sequencing are described next.

### 1.5.2 Deconvolution techniques used in DNA sequencing

Previous attempts at resolving overlapping peaks using deconvolution involved entropy maximization algorithms [10, 11], homomorphic blind deconvolution [12], and iterative deconvolution methods such as van Cittert algorithm [13, 14].

The maximum entropy method extends the model described in (1.5) to include measurement noise, $n$ [15]:

$$y = Dx + n$$  \hfill (1.7)

or expressed in terms of 1-dimensional signals:

$$y_k = \sum_{i=1}^{N} x_{i}g_{k-i} + n_k \quad (k = 1, \ldots, N)$$  \hfill (1.8)
where \( x_i \) is the i-th sample of the 1-dimensional signal \( x \), \( y_k \) is the k-th sample of the observed signal \( y \), \( g \) is the point spread function (PSF), \( n_k \) is the measurement noise at the k-th sample point and \( N \) is the length of the observed signal. The maximum entropy method chooses among a feasible set of solutions, the particular solution which is consistent with the observed data. Expressed in the form of an optimization algorithm, maximum entropy method can be stated as maximizing the following expression:

\[
a S(x, m) - C(x)
\]  

(1.9)

where \( C(x) \) is a constraint used to choose only the feasible solutions, \( S(x, m) \) is known as configurational entropy and \( a \) is a regularization parameter which is successively reduced during iteration. When the noise process is assumed to be normally distributed with a variance \( \sigma \), the constraint is given by:

\[
C(x) = \sum_{k=1}^{N} \left( \frac{\sum_{i=1}^{N} x_i y_k - y_k}{\sigma_k^2} \right)^2
\]  

(1.10)

The configurational entropy is a measure of similarity between the resolved signal at a particular iteration and a signal model, say \( m \):

\[
S(x, m) = \sum_{i=1}^{N} \left( x_i - m_i - x_i \log \frac{x_i}{m_i} \right)
\]  

(1.11)

This formulation is well suited for inclusion of prior knowledge in the form of the reference signal model \( m \). Though the authors in [11] ignored the existence of prior knowledge, the authors in [10] used a heuristic derivative-based peak finding method to arrive at the signal model from the resolved waveform. The obvious drawback of this approach is that such heuristic peak finding methods can result in spurious peaks being amplified and subsequently emphasized in successive iterations. Furthermore, the entropy maximization method requires accurate description of the PSF. Since the PSF usually varies along the length of a sequencing lane, this method has to be restricted for use over small regions of DNA time traces. As stated in [11], prior
information about mean peak separation cannot be explicitly integrated into this algorithm. Instead, heuristics [16] are used to reject spurious peaks.

In [12], homomorphic blind deconvolution is presented as a means of resolving overlapping peaks. This method is aimed at overcoming the need for exact knowledge of the PSF. The homomorphic blind deconvolution algorithm assumes that the observed signal is the result of convolving an impulse train (representing the true peaks) and a Lorentzian PSF. The log-spectrum of the Lorentzian is a straight line with a negative slope, whereas, the frequency distribution of the pulse train is almost flat. Hence, the Fourier transform of the log-spectrum (i.e. the cepstrum) exhibits a low frequency component predominantly due to the PSF. This low frequency component is removed using a high-pass filter in the cepstral domain (i.e. a high-pass lifter). Subsequent reversal of the processes is expected to remove the influence of PSF in the signal and hence sharpen the peaks in the time-domain. The authors argue that the use of a general high-pass filter that is not fine tuned to a particular PSF makes the procedure blind. The application of high-pass filter leads to loss of some energy from the impulse train and the generation of spurious peaks. The latter effect can be suppressed to certain extent using a low-pass filter which is applied to the filtered cepstrum. But the bandwidth of the low-pass filter determines the resolution of the final sharpened peaks. Results in [12] show that further processing is required to suppress the appearance of spurious peaks. If this procedure is applied indefinitely, the final results converge towards an impulse train and the amplitudes of the peaks are not preserved in this process. Like the maximum entropy method, this method also does not possess an explicit means of incorporating statistical prior knowledge about peak separation.

The iterative deconvolution approach to enhancing resolution in DNA data [13, 14, 17] employ a modification to the classical iterative scheme proposed independently by van Cittert [18] and Landweber [19]. According to the general form of the iteration, the estimate of the unknown input \( x \) at the \( k \)-th iteration is given by:

\[
x^{(k)} = x^{(k-1)} + \beta(y - Dx^{(k-1)})
\] (1.12)
in which, as before, $y$ is the observed or measured signal and $D$ is the point spread function. This is the general form of the van Cittert iterative deconvolution scheme. Intuitively, this heuristic algorithm could be explained as follows. In each iteration, the estimate of the true signal is obtained from the previous estimate by adding a fraction of an error term. The difference between the observed signal and the smeared version of the previous estimate serves as this error term.

Formulation in [20] showed that additional constraint operators can be incorporated with the distortion operator $D$ to obtain similar iterative schemes which included additional constraints. For example, if an arbitrary constraint is defined by the operator $C$, then the observed waveform can be redefined as:

$$y = DCx$$

where the constraint operator is mathematically defined as:

$$x = Cx, \quad \text{iff } x \text{ satisfies the constraint}$$

Using (1.13) in (1.12) leads to the following iterative restoration scheme:

$$x^{(k)} = Cx^{(k-1)} + \beta (y - DCx^{(k-1)})$$

The advantage of this iterative scheme is that prior knowledge such as positivity or finite support of the signal $x$ can be easily incorporated to constrain the solution space. The convergence of this algorithm has been investigated extensively in [8, 20, 21, 22, 23, 24]. Some of these analyses and new insights gained through empirical investigations are presented in Section 2.2 and Appendix A.

The modified approach used for DNA data follow from the adaptation proposed by Jansson [25], where the constant relaxation factor, $\beta$, is replaced with a heuristic signal dependent term:

$$\beta_i = 1 - \frac{2}{\gamma} \left| x_i^{(k)} - \frac{\gamma}{2} \right|$$
Here, the constant $\gamma$ is set to the maximum allowable height of isolated homozygous peaks. When the signal level at a particular sample point $i$ and iteration $k$ reaches a value of either 0 or $\gamma$, $\beta_i$ reduces to zero. Hence, no correction is applied at sample point $i$ in that iteration. For any values between 0 and $\gamma$ the relaxation factor is proportional to the deviation of the estimated signal from $\gamma$. Hence, estimated values of the signals are constrained between 0 and $\gamma$. In other words, positivity constraint and upper bound are incorporated into the iterative deconvolution scheme. Unfortunately, the signal dependence of the relaxation term makes it tedious to perform convergence analysis on this modified method. Also, the van Cittert method or its modified forms are not amenable to incorporation of prior knowledge about the statistics of the process. Only loose constraints can be incorporated.

1.5.3 Other methods

A pattern recognition approach to signal conditioning is presented in [26]. In this method, an artificial neural network (ANN) [27] is used to map five successive samples of the fluorescence data from all four lanes (A, C, G and T) to four output nodes that indicated the identity of base present in the middle of the set of samples, as shown in Figure 1.8. Each of the four output nodes correspond to one of the four base types and the output node with largest level indicates the base type present in current sample set. When no peaks are present, none of the outputs produce a strong output. The proposed ANN architecture is comprised of processing layers (known as hidden layers) between the input nodes and the output nodes with non-linear transfer functions. Sigmoid function is used as non-linear transfer function to constrain the output of each node to the range 0 to 1. Approximately 400 such nodes are used in the hidden layer. The output nodes are also assigned a non-linear transfer function. The input is multiplied by a set of weight values and then applied to the hidden layer. This procedure is repeated using another set of weights between the hidden layer and the output nodes. Hence, the final output of the network is a highly non-linear mapping of the input space to the output space. A data set with corresponding expected output values (known as a training data set) is used to adjust the weight values so that the neural network's input-output pattern matches the training set.
Generally, a gradient search method is employed to optimize the weight values so that the outputs of the ANN closely match that of the training set. The training set comprised of approximately 2000 to 3000 bases which were previously identified. The inner product of the output node values and the corresponding raw fluorescence data is used to generate the transformed data stream. It is argued that this method would preserve relative peak heights and peak separations in the transformed data. The drawback in this method is that the inner product uses the raw fluorescence data and does not account for bias introduced to a particular peak's height and position by neighbouring peaks, especially when successive peaks are present in the same lane. Hence, true peak heights and separations may not be preserved. The use of large set of parameter values (weight values) and non-linear functions at each layer of the network, makes it impractical to traverse backwards to explain a particular decision of the processor. The authors only illustrate examples where the peaks are already reasonably well separated. Also, this method requires a large set of tagged data which has been collected using similar DNA sequencing system. Since the purpose of our research is to gain understanding of the parameters of the system, this method is not suited for further exploration.

More recently, a non-linear peak fitting algorithm is developed [28]. A parametric attempt is made to model the DNA time trace as a superposition of Gaussian functions. The algorithm for Gaussian functional decomposition could be summarized as follows:

1) Utilize second derivatives to identify segments in the DNA time trace which are likely to contain a peak.

2) Sort the located segments based on heuristics that incorporate the degree of overlap with adjacent peaks.

3) Perform nonlinear fit of Gaussian peak shape to the top one-third of the identified segments.

4) Discard segments with poor fit or excessive width.

5) Subtract estimated Gaussian functions from the time trace.
6) Repeat until no segments exist above the noise level.

The above method is repeated independently for each of the four lanes. As evident from the algorithm, in the case of overlapping peaks, successive estimation and subtraction is performed to identify and remove members in an overlapping segment. When the overlapping peaks are not of the same amplitude, the accuracy with which each component is estimated can affect the successful removal of subsequent peaks. In fact, it is identified in [28] that step 2 in the above procedure critically determines the ability of algorithm to resolve overlapping peaks. Further, the simple superposition model does not incorporate explicitly any prior knowledge about the statistics of the signal (for example, mean separation of peaks). Hence, rejection criteria (step 4) need to be used to exclude those Gaussian fits which were not realistic. For the same reason the author showed that use of asymmetric peak shape description resulted in failure of the method to resolve shoulder peaks from adjacent strong peaks.

A more structured approach was developed by our group [5] where a statistical model of the DNA time trace was presented and an optimal detector was derived based on this model. The model incorporated the salient features of the DNA time trace, namely, amplitude, peak shape and peak separation. The statistical formulation facilitated easy modification of the model and the receiver to cope with new cases which were not originally incorporated in the model. Details of the model are provided in Section 2.3.

Another processor based on maximum likelihood equalization of digital communications channel was presented in [29]. In this work, the base calling is performed in two stages by separate processors: a soft caller followed by a hard caller. The time series is broken into windows of approximately 30-40 bases. The variation in peak spacing between peaks is assumed to be small within these windows. The soft caller is applied independently to each of the four lanes to identify tentative peaks corresponding to bases. For each lane the soft caller minimizes the square error between the actual DNA time-trace and an estimated trace. The estimated trace is constructed using locations of peaks in previous windows, estimated peak shape and a minimum peak spacing constraint for candidate peaks. The hard caller is applied jointly on the current window of candidate calls for all four lanes. The processor removes those ten-
tative calls which do not correspond to actual bases. This is done using a cost value generated using heights and positions of tentative calls. The *hard caller* evaluates this cost using forward dynamic programming and produces the subset of candidate calls which is the best approximation of the *ideal* set of evenly spaced, large-amplitude peaks. Though the method adopts statistical formulation, the assumption of a rather simplistic signal model limits its applicability. For example, the method does not incorporate peak height variance nor peak-spacing variance. This assumption would limit its applicability in regions where variance in peak-spacing or peak-heights is large – such as towards the end of the DNA time-series. This maximum likelihood formulation only accommodates variance introduced by electrical noise which is modeled as an additive white Gaussian noise. But, examination of DNA time series have shown that the electrical noise component is small relative to the artifacts introduced by the chemical reactions.

1.6 Detecting polymorphic sites

From the perspective of DNA time-trace, heterozygous sites and partial mutations bare close resemblance to each other. Both sites are characterized by the presence of smaller peaks in two lanes (see Figure 1.4 illustrated previously).

1.6.1 Biological significance

The gene is generally a stable unit. Sometimes, a change occurs in the sequence of a gene either due to environmental factors or pathogen attack. Such a change is known as a mutation. Mutations are broadly classified as point mutations and rearrangement mutations [30]. In point mutation, only a single nucleotide in the sequence of a gene is affected. This, in turn, may cause a change in the protein that the gene represents. In rearrangement mutation, a large region of the gene is affected. Two simple types of rearrangements are insertion of additional sequence to the gene and deletion of a segment of the gene.

Of the two types of mutations described above, substitution type mutations of single bases (substitution-type point mutations) are the most frequent form of DNA
sequence variation in the human genome [31]. Accurate identification of these mutations and heterozygous sites have been shown to be crucial for detailing the evolutionary history of human populations. From a clinical perspective, a large number of disease-related genes have been genetically mapped and the identification of these disease genes (for presymptomatic disease testing) requires accurate detection of disease causing mutations ([32, 33] are but a few examples). Also of clinical value is identification of single nucleotide substitutions for matching tissues prior to transplantation [34, 35] and for analyzing samples in forensic situations [36].

1.6.2 Current algorithms

Both commercial and academic groups have developed algorithms that automate detection of polymorphic sites. These methods attempt to circumvent some of the limitations identified in the previous section.

In [37], the authors present a method to simultaneously compare sequences from different individuals. In this method, only one (or at most two) base type is selected and the time series obtained by sequencing the selected base type from different individuals is overlapped to generate a consensus pattern. Any change in the height of a peak from a particular sample is expected to be clearly visible relative to the consensus pattern. The utility of this technique is illustrated by analyzing base substitutions in the human β-amyloid protein precursor gene. This method is primarily aimed at screening mutations in a large number of DNA samples from different individuals. The fundamental assumption of this method is that particular mutations being analyzed are known in advance and hence appropriate base types can be chosen in advance for the sequencing process. The method exploits the high degree of correlation that exists between lanes within the same gel. In the case of unknown mutations, all four lanes must be analyzed and hence this method does not provide any advantage over independent sequencing of all the samples and their subsequent comparison. Detection of heterozygous sites requires at least two lanes and hence this method is not suitable for detecting such variations. Furthermore, the presence of noise peaks can lead to ambiguities since the other three nucleotide time traces are not present to ascertain the identity of a peak (i.e. as a noise peak or a valid peak).
In [35, 38], the authors describe a heterozygote detection algorithm that depends on the peak heights of the signals. For example, if the horizontal axis is used to represent the peak heights along one lane and the vertical axis is used to represent the peak heights along another lane, the peak heights of homozygous peaks would tend to cluster along the axes. But heterozygous combinations resulting from the presence of peaks in the above mentioned lanes would cluster in the mid region of the plot. The advantage of this method over the previous method is that it incorporates information from two lanes to decide on the identity of a putative heterozygous site. The disadvantage of this method is that no attempt is made to account for bias introduced by adjacent peaks on the estimated peak heights. Though the authors demonstrate an example where the correct heterozygote is detected even when the peaks do not perfectly align with each other (see issue (c) in Section 1.4.2), they do not provide sufficient details regarding the algorithm or the performance so that it may be critically analyzed.

A method specifically designed for detecting heterozygous peaks in traces obtained using dye-terminator chemistry is described in [39]. In this method, the amplitude of each peak is normalized using average peak heights computed using two 25-base windows immediately upstream and downstream. Peaks that are off-scale, less than 10% of full-scale, or within 3 bases of the base in question are excluded from the set used for normalization. Hence this technique, unlike the previous method, accounts for bias introduced by adjacent peaks. Bases adjacent to the heterozygous site are excluded because the peak heights at these positions are observed to be depended on the sequence, particularly when dye-terminator chemistry is used. Any base location where the normalized peak height is reduced by more than 40% and another base-peak is simultaneously present in another lane is considered a heterozygous site. The authors also attempted to distinguish between false positives introduced by the presence of spurious peaks. This was achieved by the inclusion of additional heuristics which required significant change in peak height in the base located on the 3' side of the putative nucleotide variation. The authors recognized that the heuristic nature of the algorithm does not facilitate a means of optimally estimating the absolute false-negative rate of the method.
Chapter 1 • Introduction

A similar method to the one described above is presented in [40] from the perspective of HLA typing by direct sequencing. In this method, specialized chemistry (solid-phase T7 sequencing chemistry) is used to achieve even base incorporation and hence uniform peak heights. This simplifies the detection algorithm. A heterozygote is detected when the peak height is 40-50% of the height of homozygous peaks. Normalization is also avoided due to the availability of uniform peak heights. But this method fails to accommodate for bias introduced by adjacent peaks in the same lane. Although this method may be suitable for identifying heterozygous sites in high resolution region of the DNA time-trace, which is typically observed towards the beginning of a sequencing run, it will fail to accurately identify sites located towards the end of the run.

An alternative approach to comparing peak-heights is presented in [41]. Generally, a heterozygote is detected if the ratio of the two strongest peaks (say, $I_1/I_2$) at a base position exceeds a set threshold value. This methodology works well provided the traces are of high quality and the peak heights are fairly uniform. Clearly, presence of spurious peaks can result in false positives. In the proposed method, an additional peak height comparison is introduced to eliminate a large number of such false-positive heterozygote calls. The modification involves the calculation of the ratio of the second and third strongest peaks ($I_2/I_3$) and ensuring that this ratio is below another threshold. The presence of a high $I_2/I_3$ is interpreted as a hard stop. This work indirectly hints at the need to consider all four lanes when deciding on the identity of a particular base position but fails to provide a structured approach to incorporate this higher level information.

More recently, in [42], the method described above was extended to incorporate quality measures. The concept of quality measure for the identity of each base position is introduced in [43]. Though a clear computation method is not provided in the literature, the quality measure is stated to be generated using three factors: peak spacing, relative size of the uncalled and called peaks, and the decrease in signal strength between identified peaks. The authors note that the performance of the

---

2Condition where three or four peaks are present at the same base position. This is an artifact of the chemical processing rather than a biological phenomenon.
algorithm is quite closely related to the quality measure of the data considered. Hence, the performance of the proposed method would deteriorate towards the end of the sequencing run where quality values are generally poor.

All of the methods described above are developed based on heuristics formulated through experience in specific systems. None attempt to take a structured approach. A structured approach will define a signal model and then derive an appropriate detector. This approach will facilitate evaluation of the performance bounds of the detectors and hence provide a means of assessing risk associated with the detection system – a requirement for most biomedical device safety standards. As correctly pointed out in [39], no optimal strategy exists to estimate the absolute false-negative rate of the heuristic methods.

1.7 Problem statement

Given a segment of low-resolution DNA time trace with overlapping peaks, i.e. a complex, one wishes to develop a unified scheme that detects the number of peaks in the complex and the local sequence in this segment. Such a local processor would be well suited for application as a statistical proof reader for existing base-calling programs.

1.8 Research approach

Based on the statistical model developed by our group [5], a statistical processor is derived for resolving overlapping peaks that may be heterozygous or homozygous. The performance of the proposed processor is assessed from two perspectives – detection of overlapping homozygous peaks and detection of heterozygous/mutation sites. The former case is assessed using both simulated and real-life data and compared with a commonly used method. The latter case is assessed using a large set of real-life data and compared with the performance of a commercial software. In both cases, attempt is made to derive a probability measure associated with the decisions of the processor.
1.9 Dissertation organization

In this thesis, Chapter 2 further explores the iterative deconvolution scheme and its modifications. This work was done to gain insight into the problem and served as a motivation for developing a non-conventional approach to deconvolution. This chapter also describes a statistical signal model that is used to describe the DNA time series. Then a post-processor to correctly identify local sequences in regions with low confidence is derived. Chapter 3 describes the use of the proposed processor as a proof reader using real DNA data. Then the performance of the proposed processor is assessed using both simulated and real DNA data. The performance is also compared with a commonly used peak sharpening algorithm.

The ability of the proposed structure to cope with clinically significant special cases is illustrated through the use of heterozygotes and partial-mutations in Chapter 4. Here, a large set of real DNA data is used to compare the performance of the processor with that of a commercial base-calling program. Finally, in Chapter 5, summary of contributions of this work are outlined and avenues for future research are suggested.
Figure 1.8: Mapping of inputs and outputs of an ANN which serves as a signal conditioning processor (see Section 1.5.3 for description). Note that each connection (shown as lines) between processing nodes (shown as circles) has an associated weight value. The relative levels of the output nodes provide a measure of certainty associated with the respective base types.
CHAPTER 2

Detection of DNA sequences in low-resolution data

2.1 Introduction

Prior to the development of the maximum likelihood local post-processor (presented in Section 2.3), the applicability of commonly used iterative deconvolution algorithms were explored. In particular, the Jansson-van Cittert iterative deconvolution (JID) algorithm, which was generally used to sharpen peaks in DNA data, was analyzed with the aim of incorporating a priori knowledge for the detection in low resolution data. In the next section, the conditions necessary for convergence of the general iterative deconvolution algorithm are described. These explorations revealed the limitations of the JID algorithm and motivated the development of a new statistical scheme. This development is described in Section 2.3.
2.2 Analysis of van Cittert algorithm

In [44], Bracewell and Robers analyzed the following form of the general van Cittert algorithm (1.12) where the relaxation factor, $\beta$, is set to one.

$$x^{(k)} = x^{(k-1)} + (y - Dx^{(k-1)})$$  \hspace{1cm} (2.1)

This simplified iteration is represented in the frequency domain as:

$$X^{(k)} = X^{(k-1)} + [Y - GX^{(k-1)}]$$  \hspace{1cm} (2.2)

where $X$ is the Fourier transform of estimated original signal $x$, $Y$ is the Fourier transform of the observed signal $y$ and $G$ is the frequency domain representation of the point spread function (PSF). As before, the superscripts in parentheses represent iteration number. Convolution theorem is used to convert the convolution between the PSF and previous estimate of the original signal to a product of the respective frequency domain representations. For the purpose of analysis, the functional form of the PSF, $D$, will be represented by $g$. Like $x$ and $y$, the dependence of $g$ on sample time is implicit.

If the initial condition for the iterative equation in 2.2 is assumed to be:

$$X^{(1)} = Y + [Y - GY]$$  \hspace{1cm} (2.3)

then the iteration can be rewritten using successive substitution as:

$$X^{(k)} = \{1 + [1 - G] + [1 - G]^2 + \ldots + [1 - G]^k\}Y$$  \hspace{1cm} (2.4)

The first term on the right hand side of the above expression is a geometric series and it converges provided the following condition is satisfied:

$$|1 - G(s)| < 1$$  \hspace{1cm} (2.5)

Here, the frequency dependence of $G$ is explicitly written with $s$ representing fre-
quency. Hence, the necessary and sufficient conditions for convergence of (2.1) are as follows:

\[ |1 - G(s)| < 1, \quad s : G(s) \neq 0 \]  
\[ Y(s) = 0, \quad s : G(s) = 0 \]  

The conditions outlined in (2.6) are known as BracewellRoberts conditions.

In [21], the above conditions are illustrated using practical constraints that the PSF should satisfy. For example, the first part of BracewellRoberts condition is restated as \( 0 < \text{Re}[G(s)] < 2 \), where \( \text{Re}[G(s)] \) is the real part of \( G(s) \). Further, if the PSF \( g \) is real, then \( G \) is Hermitian (i.e. \( G(s) = G^*(-s) \)). Hence, the first part of BracewellRoberts condition can be written as:

\[ 0 < G_e(s) < 2 \]  

where \( G_e \) is the even component of \( G \). Since \( G \) is Hermitian, \( |G_e(s)| \leq |G(s)| \). Using the definition of Fourier transform [45] and the fact that \( |G_e(s)| \leq |G(s)| \):

\[ |G_e(s)| \leq \left| \int_{-\infty}^{\infty} g(t)e^{-j2\pi s t} dt \right| \]  
\[ \leq \int_{-\infty}^{\infty} |g(t)| dt \]  

In DNA data, \( g \) is the shape of an isolated peak and it has finite energy. Hence, \( g \) can be multiplied by a constant so that the requirement, \( G_e(s) < 2 \), is satisfied. It should be noted that in iterative deconvolution methods, the energy of \( g \) must be constrained to be unity (i.e. \( \int_{-\infty}^{\infty} |g(t)| dt = 1 \)). This ensures that the convolution between the PSF and the estimated signal \( x \) preserves the energy in the estimated signal. This requirement naturally satisfies the upper bound imposed on \( G_e \).

The effect of noise on the measured signal was shown in [46]. If the observed signal \( \hat{y} \) is corrupted by an additive noise term \( n \), then substituting \( \hat{y} = y + n \) in the
iterative deconvolution equation and taking the Fourier transform leads to:

\[ X^{(k)} = \left\{ 1 + [1 - G] + [1 - G]^2 + \ldots + [1 - G]^k \right\} \hat{Y} + N^{(k)} \]  \hspace{1cm} (2.11)

where the frequency domain representation of the measurement noise at the k-th iteration, \( N^{(k)} \), is given by:

\[ N^{(k)}(s) = \left\{ 1 + [1 - G] + [1 - G]^2 + \ldots + [1 - G]^k \right\} N(s) \]  \hspace{1cm} (2.12)

Here, \( N \) is the Fourier transform of the measurement noise \( n \). The contribution of noise would be evident when signal levels are low. In other words, when \( G(s) \approx 0 \), the noise level at the k-th iteration is given by:

\[ N^{(k)}(s) \approx \{1 + k\} N(s) \]  \hspace{1cm} (2.13)

Hence, the noise level is shown to grow linearly with iteration number \( k \). For this reason, a smoothing filter is often employed at each iteration to suppress amplification of noise peaks.

The above discussion assumes that \( \beta \) is 1 and provide conditions that need to be satisfied by the PSF. Hence, no insight is provided in terms of conditions that \( \beta \) must satisfy to guarantee convergence. An algebraic analysis is provided in [24] to identify requirements imposed on \( \beta \). In this analysis, the iterative deconvolution scheme is represented in matrix form as shown:

\[ \mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + \beta \left( \mathbf{y} - D \mathbf{x}^{(k-1)} \right) \]  \hspace{1cm} (2.14)

Here, \( \mathbf{x} \) and \( \mathbf{y} \) are the estimate of the original signal vector and observed vector, respectively, \( D \) represents the convolution matrix generated using the PSF and \( \beta \) is assumed to be a constant. This iterative expression can be rewritten as:

\[ \mathbf{x}^{(k)} = \beta \mathbf{y} + H \mathbf{x}^{(k-1)} \]  \hspace{1cm} (2.15)

where \( H = I - \beta D \) and \( I \) is the unit matrix. If the initial estimate of the original
Chapter 2 © Detection of DNA sequences in low-resolution data

signal $x^{(0)}$ is set to $\beta y$, then:

$$x^{(k)} = \beta (I + H + \ldots + H^k)y$$  \hspace{1cm} (2.16)

Let $\lambda_i (i \in \{1, 2, \ldots, n\})$ be the eigenvalues of the PSF matrix $D$. Then $(1 - \beta \lambda_1), (1 - \beta \lambda_2), \ldots, (1 - \beta \lambda_n)$ are the eigenvalues of $H$. If,

$$(1 - \beta \lambda_i)^k \rightarrow 0 \quad \text{as } k \rightarrow \infty, \quad \text{for all } i$$  \hspace{1cm} (2.17)

then $H^k \rightarrow [0]$ as $k \rightarrow \infty$. Hence, the sequence in (2.16) would converge and the iteration would tend towards the inverse operation:

$$x^{(k)} \rightarrow D^{-1}y$$  \hspace{1cm} (2.18)

The necessary and sufficient condition for the convergence in (2.17) is given by:

$$|1 - \beta \lambda_i| < 1$$  \hspace{1cm} (2.19)

If $\lambda$ is represented in complex notation as $\lambda_i = a_i + j b_i$, the above inequality can be expressed as:

$$\beta [\beta (a_i^2 + b_i^2) - 2a_i] < 0$$  \hspace{1cm} (2.20)

If the $D$ matrix is positive definite (i.e. all eigenvalues are positive and non-zero), then $\lambda_i = a_i > 0$. Hence, above inequality simplifies to the following:

$$0 < \beta < 2/\lambda_{\text{max}}$$  \hspace{1cm} (2.21)

where $\lambda_{\text{max}}$ is the largest eigenvalue of $D$, $\lambda_{\text{max}} = \max\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$. In other words, the constant relaxation factor $\beta$ can be chosen based on the largest eigenvalue of the PSF to guarantee convergence. Note that the positive definite requirement of
the $D$ matrix can be satisfied for any real $D$ using the following transformation:

$$
\hat{D} = D^T D D^T
$$

(2.22)

Schafer [20] arrived at a similar convergence condition using the Contraction Mapping Theorem of functional analysis [47].

Since the relaxation factor, $\beta$, in the JID method is signal dependent, the above analyses cannot be performed in this case. For this reason, simulated data was used to assess the characteristics of the JID method. Details pertaining to the implementation of this algorithm and the study of its convergence behaviour are summarized in Appendix A. Appendix A shows that the JID algorithm fails to resolve peaks when the ratio of peak width to peak separation exceeds 2. Hence, the method is not well suited for application in low resolution regions of the DNA data. Nonetheless, the JID algorithm is used throughout this work for comparison purposes.

The above analysis shows that the JID method provides means of incorporating loose constraints such as positivity but prior knowledge about the statistics of the signals cannot be incorporated. Also, the fundamental assumption that the underlying signal is a train of impulses results in loss of information about the amplitude of the components. Hence, conditions that are characterized by differences in amplitudes, such as heterozygous sites and partial-mutations, cannot be detected using this method. This implies that there is a clear need to resolve the complexes into components that resemble isolated peaks in the DNA data. Also, a new method which incorporates the \textit{a priori} knowledge about the DNA sequence might result in better performance. Examples of such \textit{a priori} knowledge include expected peak separation, expected heights of homozygous and heterozygous peaks, and presence of peaks uniquely in only one lane in the case of homozygous condition.

2.3 Maximum likelihood detection of local sequence

The statistical model that characterizes the \textit{local} behaviour of DNA time-series will be presented in this section. A maximum likelihood detector for a local region is
then derived using this model. Details pertaining to implementation of this detector are also presented.

### 2.3.1 Statistical Model

Figure 2.1 shows the entire time-series corresponding to a single channel in automated DNA sequencing. Three gross regions are evident in this graph. On the left side is a constant region corresponding to the time lapsed between the start of the electrophoresis process and the first arrival of fragments at the photo sensors. Immediately following this region is a large peak that almost saturates the photo sensors. This peak corresponds to excess fluorescently labeled primers which are not incorporated into any copies of the DNA fragment being sequenced. The large primer peak causes an exponentially decaying offset in the base-level of the subsequent region. The peaks corresponding to the presence of nucleotides of a specific type (A in this case) are superimposed on top of this varying base-level. The peaks of the nucleotides also exhibit an exponentially decaying trend and this is probably due to consumption of terminator nucleotides. If the data acquisition process is carried out well past the reading length of the sequence, another relatively strong peak may be observed. This peak is due to full-length fragments. These gross trends were removed by the pre-processing methods described in [48]. Since the intention of this work is to develop a local post-processor, the four lanes are also assumed to be aligned with respect to each other. Misalignment between lanes generally occur due to differences in mobility.

Figure 2.2 shows time series from all four lanes where the above mentioned gross effects have been compensated. The time-series exhibit peaks of similar shape and peaks occur exclusively in one of the four lanes at each base position. The base-line is not perfectly flat and shows the presence of noise processes. These characteristics motivated the following signal model [5]:

\[ y_{k,j} = \sum_{i=1}^{N} a_i g_{j,i} \delta_{k,x_i} + n_{k,j} \]  

(2.23)
where $y_{k,j}$ is the signal level of the waveform in lane $k$ ($k = 1, 2, 3, 4$) and normalized sample time $j$; $N$ is the number of bases in the local sequence; $a_i$ is the amplitude of the $i$-th peak; $g$ is the generic peak shape; $t_i$ is the peak-time of the $i$-th peak; $n_{k,j}$ is the noise contribution to the time-series in lane $k$ and sample index $j$; and $\delta_{k,x_i}$ ensures that a contribution to the waveform is made only when the base type $x_i$ at base position $i$ is the same as lane $k$. Details of the model parameters are described next.

Peak shape

As illustrated in Figure 2.3, the proximal (peaks near the beginning of a sequencing run) and distal peaks (peaks near the end of a sequencing run) matched each other when the width of the proximal peak was stretched to fit that of the distal peak. This
Figure 2.2: Segment of a DNA time-trace with large-scale trends removed.

relationship in the horizontal scale of peak shape was formalized as:

$$g_{j,t_i} = g_1 \left( \frac{j - t_i}{pw(t_i)} \right)$$

(2.24)

where $g_1$ is the generic peak shape at unit pulse width, $pw(t_i)$ is the peak width at peak time $t_i$.

It was suggested in [7] that electrophoresis of a pure molecule should give rise to a band that is Gaussian in shape. This general shape has been adopted in a number of base-calling algorithms [13, 11, 28]. But time-series obtained using Micro GeneBlaster showed that the peak shape was asymmetric. Hence, the generic peak shape was modeled as consisting of exponential tails on either side of a Gaussian shape. The decay factors of the two exponential components corresponded to the
Figure 2.3: Comparison of a proximal and a distal peak after the distal peak has been scaled horizontally. Both peaks have similar shape.

rising and falling edges of the peak. This is described mathematically as:

\[ g_1(t) = \begin{cases} 
\alpha_1 e^{\gamma_1(t-t_i)}, & t < t_i \\
\alpha_2 e^{-(t-t_i)^2),} & t_i < t < t_r \\
\alpha_3 e^{-\gamma_2(t-t_r)}, & t > t_r 
\end{cases} \]  

(2.25)

where, \( \alpha_1, \alpha_2 \) and \( \alpha_3 \) correspond to the relative amplitudes of the three components; \( \gamma_1 \) and \( \gamma_2 \) are the exponential factors; \( t_i \) is the position of the i-th base; \( t_i \) and \( t_r \) are the points that demarcate the boundary between the central Gaussian component and the two exponential components. All the parameters of the generic peak shape can be dynamically estimated from a known isolated peak in a DNA time-series. Different peak parameters are estimated for different lanes in the same DNA time-trace. Figure 2.4 justifies the choice of this generic peak shape by comparing it with
real data. Notice that the difference between the synthesized waveform and observed waveform along the falling edge of the peak is primarily due to noise.

Figure 2.4: Comparison of an isolated DNA peak and the peak synthesized using (2.25).

Peak amplitude

Figure 2.5 shows a scatter plot of peak amplitudes in the A lane of a typical sequencing run. The amplitudes of isolated A peaks were estimated after removal of the gross trends described in the beginning of this section. As part of the trend removal process, the signals were also normalized such that mean amplitude of isolated peaks was approximately unity. The amplitude estimation was carried out assuming that the sequence was known in advance. Though the range of values taken by the peak heights remain almost the same throughout the sequence, some large variations were observed towards the end of the sequencing run. These were due to poor estimation of peak heights, which resulted from poor signal-to-noise ratio (SNR). The covariance of
the peak amplitudes were estimated for lags up to 20 bases and Figure 2.6 shows the plot of covariance as a function of lags. A single sharp peak at zero lag was present and covariance values took on small values for non-zero lags. Hence, the normalized amplitude values\(^1\) can be considered to be uncorrelated. This was similar to the result observed previously by our group using ALF\(^{TM}\) sequencer\(^2\) [49]. If the covariances are assumed to be zero for all lags other than zero, the amplitudes of each peak can be modeled as an independent Gaussian random variable with unit mean and variance \(\sigma_a^2\). The variance of peak-heights in different lanes may be different. Since this work proposes to develop a local post-processor, the variance of the peak amplitude can be estimated using the amplitudes of isolated peaks which can be identified with high confidence and are located in the vicinity of the complex under analysis.

\(^1\)Data was obtained using Micro GeneBlaster by Visible Genetics Inc., Toronto

\(^2\)Pharmacia, Sweden
Figure 2.6: Peak amplitude covariance.

Peak width

Figure 2.7 illustrates a scatter plot of estimated peak widths in the A lane as a function of base number. With the exception of few estimated widths past base position 300, this graph shows that over a small window of approximately 40 bases or so, the curve can be approximated using a piecewise linear function. This is in agreement with expression 1.4 (in Chapter 1) which shows that, for small base numbers, the peak width grows linearly with respect to base number. In the particular example illustrated, isolated A peaks were absent beyond base position 300. This resulted in poor estimates of peak widths in this region and account for the large variations seen in Figure 2.7 beyond base position 300. In order to limit the number of parameters in the signal model, the width was represented using a deterministic
variable which was linearly related to the position of a peak as shown:

\[ pw(t_i) = a + bt_i \]  \hspace{1cm} (2.26)

where \( pw(t_i) \) is the width of the \( i \)-th base and \( t_i \) is its corresponding position in time domain. The model parameters \( a \) and \( b \) in a particular local region can be easily estimated using a set of isolated peaks which have been identified with high confidence.

**Noise process**

The noise in the DNA time series originate due to both physical phenomena and chemical phenomena. Examples of physical phenomena include sensor shot-noise and pre-amplifier thermal noise. Examples of chemical phenomena include non-specificity
of terminator molecules (see Section 1.3) and lack of fidelity in the polymerase chain reaction (PCR) process [50] which is used to replicate the DNA sample. While the physical component result in uncorrelated random noise, non-specificity of the terminator molecules cause errors in labeling of some fragments. Due to the random nature of chemical reactions, a fraction of the DNA fragments are wrongly labeled. This manifests as noise peaks that are similar in shape to signal peaks but smaller in amplitude and appear at expected peak-times. Exploration of data obtained from Micro GeneBlaster showed that the coloured noise component resulting from the chemical reactions was predominant over the electrical component. Figure 2.8 illustrates a segment of DNA time series where the coloured noise component is clearly visible. Though the noise shape was similar to the signal peak, Figure 2.9 shows that the amplitude of these noise peaks were observed to be uncorrelated beyond one base position within the same lane. The correlation within one base position is due to the extension of the support region of each noise peak into the support region of the adjacent noise peak. Hence, in this work the noise was assumed to be composed of peaks with shapes similar to that of signal peaks and the amplitudes of these peaks were modeled as uncorrelated Gaussian random variables. The position of the noise peaks were also assumed to coincide with the expected peak-times. As in the case of peak amplitudes, the statistics of the noise in a particular lane can be estimated by sampling the time-trace in regions where: (i) peaks are absent in the particular lane and (ii) influence from adjacent peaks in the same lane is minimal.

**Peak time**

In the L-DNA processor, the model used to describe the peak-time vector is:

$$t_i = i\mu_t + \phi_i$$  \hspace{1cm} (2.27)

where $t_i$ is the peak-time of base (i.e. nucleotide) at position $i$ in the local sequence, $\mu_t$ is the mean separation between adjacent peaks and $\phi_i$ represents the random component (referred to as jitter in communication literature) of the peak-time corresponding to position $i$. Data collected using Micro GeneBlaster showed that the percent jitter
Figure 2.8: Illustration of presence of coloured noise. The T lane with sparse peaks (and hence prominent noise-only regions) is highlighted. Notice that the appearance of small noise peaks coincide with expected peak-times (indicated by the presence of peaks in other three lanes).

(defined as the ratio of standard deviation of peak-times to mean peak separation) was less than 18% when a window of approximately 40 bases was considered (see Figure 2.10). It should be noted that the jitter was observed to be typically 35% in our previous work using ALF sequencers. Hence, a simpler model was sufficient to track the variability in the peak-times over a small window of bases. In this work, $\phi_i$ were assumed to be independent zero-mean Gaussian random variables with standard deviation $\sigma_i$ (i.e. $\phi_i \sim N(0, \sigma_i^2)$). The mean peak separation can be estimated using previously identified peaks and $\sigma_i$ can be estimated from the residual left after subtracting the expected positions from the measured positions of known peaks.
Figure 2.9: Covariance of noise peak amplitudes.

2.3.2 Local DNA Processor (L-DNA)

The Local DNA processor, L-DNA, applies a modified Maximum Likelihood [51] approach to a segment of a DNA time series containing the observed complex. The L-DNA, uses the DNA signal model described in [49] to determine the most probable local sequence, $\hat{S}$, in this DNA segment. Since the signal model incorporates peak amplitudes, $a$, and peak-times, $t$, as parameters of the system, one approach to implementing a Maximum Likelihood processor is to estimate both the amplitude and peak-time vectors and then use these estimates to calculate the likelihood of each hypothesized local sequence. When the peaks in the same lane overlap, it is very difficult to simultaneously estimate the peak amplitudes and peak-times. Instead of simultaneously estimating all the model parameters, the L-DNA processor treats the peak-times as part of the hypothesis and then integrates the hypothesis over all possible peak-times. Each hypothesis is weighted by the a priori probability of its
Figure 2.10: Peak-time jitter computed over a moving window of 40 bases. Percent jitter is defined as the ratio of standard deviation of peak-time and mean peak separation.

corresponding peak-times. For each hypothesized set of peak-times, the processor estimates a set of peak amplitudes that yields the best match to the observed waveform. These are then used in the likelihood evaluation.

The signal levels of the hypothesized $N$ peaks (bases) in the local DNA time series and the noise levels in the other 3 lanes, at each of the hypothesized base positions, are represented by $l_{k,i}$. Here, $k = \{1, 2, 3, 4\}$ corresponds to the four lanes (A, C, G and T respectively) and $i = \{1, \ldots, N\}$ corresponds to the $N$ base positions in the local sequence. Note that each hypothesized local sequence contains the same number, $N$, of bases. At each base position (each peak-time in particular), the signal level in the lane with the hypothesized peak and the noise levels in the other three lanes are used to evaluate the likelihood of the hypothesized sequence. Thus, information from all four lanes are jointly considered by the processor when testing for the presence of a
peak (base type) in a single lane (homozygous type) [52].

For a particular hypothesized sequence, \( \tilde{S} \), and hypothesized peak-time vector, \( \tilde{t} \), if \( \tilde{l} \) is the estimated signal and noise levels vector, then the most likely local sequence in the region under consideration is obtained by maximizing the posterior probability density function (pdf), \( p(\tilde{S}, \tilde{t}|\tilde{l}) \). Hence, the ML problem is:

\[
\tilde{S} = \arg \max_{(\tilde{S}, \tilde{t})} p(\tilde{S}, \tilde{t}|\tilde{l})
\]  

(2.28)

where, \( \tilde{S} \) is the most probable estimate of the local sequence. Bayes' rule provides a way to convert the posterior pdf in (2.28) into a form that is easily calculated provided a signal model is available:

\[
\tilde{S} = \arg \max_{(\tilde{S}, \tilde{t})} \frac{p(\tilde{l}, \tilde{S}, \tilde{t})}{p(\tilde{l})} p(\tilde{S}, \tilde{t})
\]  

(2.29)

where \( p(\tilde{l}) \) is the joint probability density function of the signal and noise levels. Since the proposed method assumes the same number of bases in all the hypotheses and the local sequence is assumed to be independent of the mean separation of peaks [49], the joint pdf term, \( p(\tilde{S}, \tilde{t}) \), in (2.29) can be written as the product of the pdf's, \( p(\tilde{S}) \) and \( p(\tilde{t}) \).

\[
\tilde{S} = \arg \max_{(\tilde{S}, \tilde{t})} \frac{p(\tilde{l}, \tilde{S}, \tilde{t})}{p(\tilde{l})} p(\tilde{l}|\tilde{S}, \tilde{t}) p(\tilde{t})
\]  

(2.30)

Equation (2.30) can be further simplified by observing that \( p(\tilde{l}) \) and \( p(\tilde{S}) \) are independent of the hypothesized sequence\(^3\). Hence, the L-DNA processor can be written as:

\[
\tilde{S} = \arg \max_{(\tilde{S}, \tilde{t})} p(\tilde{l}|\tilde{S}, \tilde{t}) p(\tilde{t})
\]  

(2.31)

\(^3\)\( p(\tilde{l}) \) is independent of the hypotheses because the number of signal and noise terms in the \( \tilde{l} \) vector remain the same. \( p(\tilde{S}) \) is independent of the hypotheses since no \textit{a priori} information is available about the likelihood of any particular local sequence and therefore all hypotheses are presumed to be equally likely.
Chapter 2  Detection of DNA sequences in low-resolution data

Since the signal and noise levels are independent across the four lanes and the bias from adjacent bases in the same lane can be removed using the inter-symbol interference (ISI) removal technique described in Section 2.3.3, the expression in (2.31) can be written as:

\[
\hat{S} = \arg \max_{\{S, \bar{z}\}} \prod_{i=1}^{N} \prod_{k=1}^{4} p(a_{k,i} | \hat{S}, \bar{z}) p(n_{k,i} | \hat{S}, \bar{z}) p(\bar{z}) \Theta_{k,i, \hat{S}}
\]  \hspace{1cm} (2.32)

where, \(a_{k,i}\) and \(n_{k,i}\) are the estimated signal and noise levels and the \(\Theta\) function is defined as:

\[
\Theta_{k,i, \hat{S}} = \begin{cases} 
\frac{1}{p(n_{k,i} | \hat{S}, \bar{z}) p(\bar{z})} & \text{if hypothesis includes a base at lane } k \text{ and position } i, \\
\frac{1}{p(a_{k,i} | \hat{S}, \bar{z}) p(\bar{z})} & \text{if hypothesis does not include a base at lane } k \text{ and position } i.
\end{cases}
\]  \hspace{1cm} (2.33)

The \(\Theta\) function ensures that only bases in the hypothesized sequence and the corresponding noise levels in the other three lanes will contribute to the joint likelihood calculation.

When peaks overlap, it is difficult to accurately estimate the peak-time vector. Therefore, the dependence of the estimated peak and noise amplitudes on the accuracy of determining the peak-time vector can be minimized by integrating the conditional distributions, \(p(a_{k,i} | \hat{S}, \bar{z})\) and \(p(n_{k,i} | \hat{S}, \bar{z})\), with respect to the peak-time vector:

\[
\hat{S} = \arg \max_{\{S, \bar{z}\}} \prod_{i=1}^{N} \prod_{k=1}^{4} \int_{\bar{z}} p(a_{k,i} | \hat{S}, \bar{z}) p(n_{k,i} | \hat{S}, \bar{z}) p(\bar{z}) \Theta_{k,i, \hat{S}} d\bar{z}
\]  \hspace{1cm} (2.34)

Hence, the likelihood associated with each hypothesis is evaluated by weighting the conditional probabilities of the estimated amplitudes by the corresponding confidence associated with the peak positions. The processor in Figure 2.11 illustrates this algorithm for a particular hypothesized sequence.

The processor presented above incorporates \textit{a priori} statistical information regarding peak amplitude, peak-time, noise in the DNA time series and the expected structure of the data (i.e. homozygous condition). Issues specific to the formulation
Figure 2.11: The L-DNA method. The particular hypothesized sequence with the highest likelihood is selected by the processor.
Chapter 2  Detection of DNA sequences in low-resolution data

of the conditional density functions of the peak amplitudes, \( p(\alpha_{k,i}|\tilde{S}, \tilde{t}) \), noise levels, \( p(n_{k,i}|\tilde{S}, \tilde{t}) \) and the joint distribution of the peak-times, \( p(\tilde{t}) \), are described below. These density functions are used in conjunction with (2.34) to estimate the most likely local sequence.

2.3.3 Implementation

This section deals with the implementation of the likelihood function derived in (2.34). Here, each of the terms in (2.34) are derived. In this description, the term \( \text{peak position} \) is used to describe the location of a base in the DNA sequence and \( \text{peak-time} \) is used to refer to the location of a peak (corresponding to a base) in the time domain.

Calculation of the conditional amplitude density function

The amplitude levels \( a \) are easily estimated where hypothesized local sequence does not contain adjacent peaks of the same type. But when adjacent bases of the same type are present, the peaks overlap and the amplitudes of the constituent peaks are difficult to estimate. This limitation may be overcome using waveform comparison as follows. First the local sequence \( (\tilde{S}) \) and the corresponding peak-times \( (\tilde{t}) \) are hypothesized. The components of the peak amplitude vector, \( a \), in lane \( k \) are estimated by minimizing the mean-square-error between the observed waveform, \( y_k \), and the hypothesized waveform, \( \hat{y}_k \):

\[
a = \arg \min_{\{a\}} \|y_k - \hat{y}_k\|_2
\]

where \( \|...\|_2 \) is the 2-norm operator (mean square). This is repeated for each combination of hypothesized peak-times, \( \tilde{t} \). In this thesis, the minimization is done using an implementation of Nelder-Mead simplex search algorithm [53] in the MATLAB\(^4\) environment. The hypothesized waveform, \( \hat{y}_k = [\hat{y}_{k,j}] \), is synthesized using the following

\[^4\text{The Math Works, MA, USA. http://www.mathworks.com}\]
model [5]:

\[ \hat{y}_{k,j} = \sum_{i=1}^{N} \tilde{a}_i g_{j,t_i} \delta_{k,x_i} \]  

(2.36)

where \( \hat{y}_{k,j} \) is the signal level of synthesized waveform in lane \( k \) and normalized sample time \( j \); \( N \) is the number of bases in the hypothesized local sequence; \( \tilde{a}_i \) is the amplitude of the \( i \)-th peak; \( g \) is the generic peak shape; \( t_i \) is the peak-time of the \( i \)-th peak; and \( \delta_{k,x_i} \) ensures that a contribution to the waveform is made only when the hypothesized base type \( x_i \) corresponds to the current lane, \( k \). Substituting (2.23) in (2.35):

\[ a = \arg \min_{\tilde{a}} \sum_j (y_{k,j} - \sum_{i=1}^{N} \tilde{a}_i g_{j,t_i} \delta_{k,x_i})^2 \]  

(2.37)

where \( [y_{k,j}] = y_k \) is the observed time series in lane \( k \). The above equation can be reformulated as:

\[ a = \arg \min_{\tilde{a}} \sum_j (y_{k,j} - \tilde{a}_m g_{j,t_m} \delta_{k,x_m} - \sum_{i=1,i\neq m}^{N} \tilde{a}_i g_{j,t_i} \delta_{k,x_i})^2 \]  

(2.38)

The second term in (2.38) describes the waveform associated with the \( m \)-th base. The subtraction of the third term represents the removal of inter-symbol interference from all the other peaks. If the bases are well separated, the third term will have negligible contribution in the support region of the \( m \)-th base. When adjacent bases overlap (inter-symbol interference, ISI), the third term will attempt to minimize the bias due to interference from adjacent bases.

If ISI is successfully removed, the amplitudes of the peaks in the sequence are independent and the conditional density of the estimated amplitude vector can be
Chapter 2  Detection of DNA sequences in low-resolution data

written as:

$$p(\mathbf{a}|\mathbf{\hat{S}}, \mathbf{t}) = \prod_{i=1}^{N} p(a_i|\mathbf{\hat{S}}, \mathbf{t}) \quad (2.39)$$

where $a_i$ is the estimated amplitude of the $i$-th base of the hypothesized local sequence for a particular combination of peak-times (described in Section 2.3.3). As the amplitude of a peak in each lane is normally distributed with mean $\mu_{k,a}$ and standard deviation $\sigma_{k,a}$, the pdf of the amplitude of each of the bases in lane $k$ can be written as:

$$p(a_{k,i}|\mathbf{\hat{S}}, \mathbf{t}) = \frac{1}{\sqrt{2\pi(\sigma_{k,a})^2}} e^{-\frac{(a_{k,i}-\mu_{k,a})^2}{2(\sigma_{k,a})^2}} \quad (2.40)$$

The mean and standard deviation of the amplitude probability density function for each lane are estimated from known peaks (for each lane) in the vicinity of the observed complex. For each combination of peak-times, the process described by the third term in (2.38) ensures that the amplitude estimates are independent.

**Calculation of the conditional noise distribution**

The noise level, $n_{k,i}$, in lane $k$ and base position $i$ is estimated after removal of the expected interference from the hypothesized bases, $\mathbf{\hat{S}}$, in the same lane. As any ISI from adjacent noise samples is presumed to be small, these noise levels are assumed to be independent and normally distributed. The conditional pdf of the noise levels are given by:

$$p(n_{k,i}|\mathbf{\hat{S}}, \mathbf{t}) = \frac{1}{\sqrt{2\pi(\sigma_{n}^k)^2}} e^{-\frac{(n_{k,i}-\mu_{n}^k)^2}{2(\sigma_{n}^k)^2}} \quad (2.41)$$

where $\mu_{n}^k$ and $\sigma_{n}^k$ are the noise statistics, and $n_{k,i}$ is the estimated noise level in lane $k$ at base position $i$. The mean and standard deviation of the noise pdf are estimated from noise samples (e.g. where signal peaks are not present) in each lane. The noise samples are taken at times corresponding to positions of known bases in one of the other lanes after removal of interference from known bases in the same lane. The
sampling region is restricted to the vicinity of the observed complex.

**Calculation of the joint pdf of the peak-time vector**

In the peak-time model presented in (2.27), $\phi_i$ are assumed to be independent. Hence, the joint pdf of the peak-time vector is given by:

$$ p(t) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi \sigma_i^2}} e^{-\frac{(t_i - \mu_i)^2}{2\sigma_i^2}} \quad (2.42) $$

As described previously, $\mu_i$ is the expected peak separation. The mean peak separation between adjacent peaks was estimated using a small window of known bases in the vicinity of the complex. This estimate was used to remove the peak-separation dependent term in the peak-time model. Then the standard deviation, $\sigma_t$, was estimated using the residual of this operation.

Our L-DNA implementation forms a discrete approximation to the integral in (2.34). Each element of the continuous peak-time vector is quantized into $M + 1$ discrete points\(^5\) over the regions $i\mu_t \pm 2.5\sigma_t$. The area under each probability density function is normalized to one, accounting for the truncation of the Gaussian function.

The quantized peak-time values for each of the $N$ bases in the hypothesized sequence are represented by the corresponding row of the $N \times (M + 1)$ matrix, $R$, described below:

$$ R = \begin{bmatrix} \mu_t - \frac{M}{2} \Delta t & \ldots & \mu_t & \ldots & \mu_t + \frac{M}{2} \Delta t \\ 2\mu_t - \frac{M}{2} \Delta t & \ldots & 2\mu_t & \ldots & 2\mu_t + \frac{M}{2} \Delta t \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ N\mu_t - \frac{M}{2} \Delta t & \ldots & N\mu_t & \ldots & N\mu_t + \frac{M}{2} \Delta t \end{bmatrix} \quad (2.43) $$

Here, the incremental peak-time interval, $\Delta t$, is given by:

$$ \Delta t = \frac{4\sigma_t}{M + 1} \quad (2.44) $$

\(^5\)Where $M$ is even.
Chapter 2  Detection of DNA sequences in low-resolution data

Evaluation of the likelihood associated with each hypothesis

Using the quantization of the peak-time vector, described previously, and rectangular approximation to integration, the expression in (2.34) can be written as:

\[ \hat{S} = \arg \max_{\tilde{S}} \prod_{i=1}^{N} \prod_{k=1}^{4} \left( \sum_{j=1}^{M+1} \cdots \sum_{q=1}^{M+1} \sum_{r=1}^{M+1} p(a_{k,i} | \tilde{S}, \tilde{\tau}) p(n_{k,i} | \tilde{S}, \tilde{\tau}) \Theta_{k,i,\tilde{S}} \right) \]

\[ p(t_{1,j}) \cdots p(t_{i,q}) \cdots p(t_{N,r}) (\Delta t)^N \]

where \( p(t_{i,q}) \Delta t \) is the probability that the peak-time of the \( l \)-th base is in the range \( t_{i,q} \pm \frac{\Delta t}{2} \). Notice that the value of \( t_{i,q} \) corresponds to the \( l \)-th row and \( q \)-th column of the \( R \) matrix; each peak-time vector is given by \( \tilde{\tau} = [t_{1,j} \ldots t_{l,q} \ldots t_{N,r}] \) and in the calculations presented in Equation (2.45), \((M+1)^N\) different peak-time vectors, \( \tilde{\tau} \), are considered.

Using (2.40), (2.41) and (2.42) in Equation (2.45), the likelihood associated with each of the hypothesized local sequences, \( \hat{S} \), is evaluated. The processor selects the sequence, \( \hat{S} \), with the highest likelihood.
CHAPTER 3

Performance Analysis

3.1 Introduction

Initial analysis using DNA time series showed that two parameters significantly affected the performance of the L-DNA processor: a) the extent of overlap between adjacent peaks (resolution ratio) and b) the uncertainty in the position of the peaks (percent jitter). The resolution ratio is defined here as the ratio of full-width at half-maximum (FWHM) of the peaks to their mean separation. The percent jitter is defined as the ratio of standard deviation of peak-time to the mean peak separation. Simulations were used to quantify the sensitivity of the processor to these two parameters. Real DNA data from a DNA sequencer (Micro GeneBlaster sequencer, Visible Genetics Inc., Toronto, Canada) were used to demonstrate the performance of the L-DNA processor as a proof reader and as a viable method for extending the length of DNA segment that can be read in one sequencing experiment.

As indicated in Chapter 1, a commonly used method for resolving overlapping peaks is the Jansson-van Citter iterative deconvolution (JID) algorithm. The ability of this algorithm to resolve overlapping peaks with that of the L-DNA processor was
also compared. Steps taken to optimize the performance of the JID algorithm and guarantee convergence of the iterative algorithm are explained in Appendix A.

3.2 Analysis using simulation

3.2.1 Implementation

Simulated DNA time series was generated using the model described in Section 2.3.1. The simulated data consisted of four time series, one per lane. Each time series was composed of 10 bases, i.e. a total of 40 bases in the sequence. In a randomized fashion a complex of up to 3 peaks was introduced in the middle of one of the lanes. The probability that the complex would consist of 1, 2 or 3 peaks was the same (1/3). For the simulation, the DNA model used a generic Gaussian peak shape, a linear model for the peak widths and Gaussian random variables to describe the amplitudes, peak-times and noise levels of the time series. Unlike the model of [5], the simulation assumed percent jitter to be uncorrelated. The mean and standard deviation of the random variables reflect values observed in low resolution regions of real data – as is the case towards the end of the DNA time series. For all simulated data, the expected value of the peak amplitudes in each lane was set to unity and the expected value of the noise levels was set at 0.1. The standard deviations of peak amplitudes and noise levels were set to 3% of the mean peak and noise amplitudes, respectively. The number of peaks in a complex was limited to 3, though it can be easily extended to more peaks. Figure 3.1 compares a segment of DNA data with the waveform that was synthesized using the signal model described in 2.3. The parameters of the signal model were estimated from peaks in the vicinity of the illustrated segment. It is evident that the synthesized waveform closely resembles the original segment within statistical variations.

To emulate the scenario during the processing of the real data, the statistics of the synthesized time series were not assumed to be known a priori. Therefore, the parameters of the signal model were estimated from the known peaks and noise segments in each lane. The performance of the detector was characterized by calculating
Figure 3.1: Comparison of DNA time-trace and synthesized time-trace. (a) DNA time-trace. The local sequence consists of a complex formed by the overlap of two G peaks. Hence the local sequence in this case is CGGA (b) time-trace synthesized using the statistics estimated from the DNA data.

the probability of error, defined as the percentage of trials where the local sequence with the highest likelihood does not match the synthesized sequence. For comparison, the probability of error of the JID processor was also calculated. The JID method was used with the same signal shape as that used for the L-DNA processor and the number of iterations was limited to 50 in order to suppress amplification of spurious noise. The number of peaks resolved by the JID processor was determined by the number of peaks that crossed a set threshold level. Details with regard to the implementation of the JID algorithm are described in Section 1.5.2 and Appendix A.
3.2.2 Results

Figure 3.2 illustrates the performance of the L-DNA processor as a function of the resolution ratio, \( R \). For each point in the graph 1000 synthesized time series were used. The percent error was calculated for resolution ratios from 0.53, corresponding to high resolution regions observed in the early part of DNA time series, to 4.8, which is well beyond the resolution ratio of 1.8 that is typically observed towards the end of DNA time series. In Figure 3.2 the local percent jitter, \( J_t \), (over a small window of approximately 10 bases in the same lane) was fixed at 14%. This is similar to that observed towards the end of a DNA time series. Two regimes can be observed in the performance curve of the L-DNA processor. In the first regime, the errors are primarily due to inaccuracies in the estimation of the parameters of the signal model and hence the error remains independent of resolution. This extends up to a resolution ratio of 3.75. In the second regime, the performance is affected by the overlap of the signals and the error increases significantly. In the JID method, the probability of error is similar to that of the L-DNA processor until a resolution ratio of 0.75 and then it increases significantly and exceeds 50% for a resolution ratio of 2.

Figure 3.3 illustrates the performance of the L-DNA processor as a function of percent jitter, \( J_t \). The simulation was performed using 10 sets of 1000 synthesized time series for each point in the graph. The 10 sets were used to estimate the standard deviation of the calculated probability of error. The performance was evaluated for local percent jitters from 14%, which is approximately the percent jitter observed in DNA time series towards the end of a sequencing run, to 28.75%. The resolution ratio was fixed at 2, which is approximately the value observed towards the end of a run in DNA time series from the GeneBlaster sequencer. Figure 3.3 shows that the probability of error increases dramatically for percent jitters greater than 22%. Since the resolution ratio for all trials was 2, the probability of error for the JID processor was above 50% (see Figure 3.2).
3.3 The L-DNA processor as a proof reader

Several segments of DNA time series (from the GeneBlaster sequencer) with complexes were analyzed. The time series obtained by sequencing a segment of the human HLA-B gene for MHC Class-1 antigen [38], using the GeneBlaster sequencer, are used to illustrate the utility of the L-DNA processor as a proof reader. The section of data consisted of a complex formed by three overlapping C peaks. As confirmed by other experiments, the sequence which incorporates the C triplet is $GGAGGCGGCCCGTG$. This was consistent with the local consensus sequence (verified using the HLA database located at http://www.ebi.ac.uk/imgt/hla/). Here, the start of sequence corresponds to base position 240 and the C triplet is present at base positions 248, 249 and 250. A small segment of the time series is illustrated in the top half of Figure 3.4. DNA time series were normalized and parameters of the signal model were estimated using the procedures described in [5]. The signal param-
Figure 3.3: Sensitivity of the L-DNA processor to uncertainty in peak-time (jitter). $\sigma_t$ is the standard deviation of the peak-time described in (2.27). The estimated standard deviation of the probability of error is also shown.

Parameters for each lane were estimated using approximately 10 bases per lane, which were known with high confidence from the vicinity of the complex.

The lower half of the figure shows the original complex observed in the C lane and the results obtained using the JID and L-DNA processors. The JID method was used with 50 iterations and the same peak shape as that used in the L-DNA processor. The number of iterations was limited to suppress amplification of spurious noise. For the L-DNA processor, the hypothesized local sequence with the highest likelihood was $CCC$, while the next competing hypothesis was $CGC$ (all other hypothesized sequences had much lower likelihoods). The confidence associated with the decision to select $CCC$ was estimated as follows. Based on the statistics of the real data, 1000 realizations of the $CGC$ time series were generated. When these synthesized time series were tested by the L-DNA processor, $CCC$ was never selected as the hypothesis with the highest likelihood. This suggests that for the statistics observed
Figure 3.4: Comparison of the JID and L-DNA methods using DNA time series. Top: Segment of the DNA time series with complex at approximately 2700. Bottom: The observed signal or the complex (solid line), deconvolved signal generated by JID method (dash dot) and the three estimated components obtained from L-DNA method, corresponding to the hypothesis CCC (dashed lines).
in this example, the probability that the L-DNA processor will select *CCC* when *CGC* is the true sequence is less than 0.1%. The constituent peaks generated by the L-DNA method (shown in lower half of Figure 3.4) were estimated using the mean value of the peak-time positions of each of the elements in the *complex* and the corresponding optimized amplitudes.

The data illustrated in the above example had an approximate percent jitter of 14% and a resolution ratio of 1.5 (corresponding to a FWHM of approximately 11 time samples and a mean separation of 7.5). As evident from the simulation results that are presented in Figures 3.2 and 3.3 the JID technique is expected to fail to resolve *complexes* with this resolution ratio, while the probability of error of the L-DNA processor is expected to be below 0.1%.

### 3.4 The L-DNA processor as a sequential decoder

Loss in resolution causes most automated base-calling algorithms to have large probability of error towards the end of a sequencing run [26, 54, 55]. Since the L-DNA processor is robust over a wide range of resolution ratios (see Figure 3.2), the L-DNA processor can be used as a sequential decoder to extend the read length\(^1\). When the L-DNA processor is used as a sequential decoder, peaks that are identified with high confidence can be used iteratively to identify subsequent bases. Hence, this method can be applied in a sequential fashion to extend the read length \(N\) bases at-a-time.

It should be noted that when L-DNA is used as a proof reader, the signal model parameters could be estimated from bases on either side of the *complex*. Using the known peaks on either side of the *complex*, the influence of neighbouring peaks could be subtracted to minimize ISI. In the application as a sequential processor, the signal model parameters can be obtained only from the bases preceding the *complex* and interference from peaks located beyond the *complex* can only be dealt with on a statistical basis. The hypothesis space was extended to \(N+1\) bases so that ISI from the base located immediately past the *complex* could be removed.

\(^1\)The length of DNA that can be read in one sequencing experiment.
A segment of DNA time series that was processed by the L-DNA processor in a sequential manner is illustrated in Figure 3.5. This data was obtained by sequencing a segment of the human HLA-A gene for MHC Class-1 antigen [30]. Other experiments confirmed that the local sequence in this segment (from base positions 230 to 253) is GGGACACGGATGTGAAGAAATACC. In this example, the read length was extended 3 bases at-a-time. The bases were known with high confidence up to base position 243. This was in agreement with the consensus sequence in the HLA database. The resolution ratio and percent jitter of this segment of the data were approximately 1.6 and 12% respectively. The base types indicated in Figure 3.5 beyond base 243 are the bases that belong to the hypotheses with the highest likelihood. Table 3.1 summarizes the sequences with the highest likelihood and the corresponding competing hypotheses for each of the 3 local bases.

The confidence associated with the decisions of the L-DNA processor was eval-
Table 3.1: The L-DNA processor as a sequential decoder: sequence of most likely triplet of bases and the nearest competing hypotheses.

<table>
<thead>
<tr>
<th>Base position</th>
<th>Most likely triplet</th>
<th>Competing Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>244..246</td>
<td>A A G</td>
<td>A A A</td>
</tr>
<tr>
<td>247..249</td>
<td>A A A</td>
<td>A A T</td>
</tr>
<tr>
<td>250..252</td>
<td>T A C</td>
<td>T C A</td>
</tr>
<tr>
<td>252..254</td>
<td>C C T</td>
<td>C C C</td>
</tr>
</tbody>
</table>

The simulation results show that the performance of the L-DNA processor is comparable to that of the JID processor when the resolution ratio, $R$, is below 0.75. Resolution ratios that are lower than 0.75 occur in the early part of DNA time series where the bases are well resolved. In regions of the time series where the resolution ratio is greater than unity and less than 2, the probability of error of the L-DNA processor is less than 0.001, whereas, the probability of error of JID processor is at least an order of magnitude larger. Since the resolution ratios of data towards the end of the DNA time series is approximately 2, the L-DNA processor will provide a significant improvement in performance for current DNA sequencers that use iterative deconvolution techniques. The significant improvement in performance of the L-DNA processor over JID processor was verified using data from a DNA sequencer (GeneBlaster).

When the L-DNA processor is used as a sequential decoder, the processor can be repeatedly applied until the resolution ratio approaches 3.75 without significantly...
increasing the probability of error (as illustrated in Figure 3.2). When the processor is used in this manner, model parameters are continuously updated by using data from bases which were previously identified by the processor. If the peak-width model is assumed to remain linear, the proposed sequential decoder could be theoretically used to increase the read length by about 90% beyond the present limit. In reality, however, the onset of nonlinearities and deterioration in accuracy of the estimated parameters could limit the increase in the read length. This is evident in the example illustrated in Section 3.4. In this example, the probability of false detection is approximately 2% compared to the expected probability of error of almost 0.1% in the simulation. Notice that in the simulation the bases are assumed to be known on both sides of the complex, whereas, in the application of the L-DNA processor as a sequential decoder the bases are known only on one side of the complex. Hence, bias introduced by the base located beyond the complex cannot be accurately removed in the latter case.

The robustness of the L-DNA processor is derived from: a) the ability to accurately estimate the local statistics of the DNA time series, and b) the ability to use the pdf's of model parameters as weighting functions to represent the certainty with which the model parameter values are known. Apart from the above two reasons, the significant difference in performance between the L-DNA processor and the JID processor is due to the fact that the L-DNA processor uses information from all four lanes of the DNA time-series, while the JID processor uses only one lane. Should the technology or the chemistry behind the electrophoresis process change and introduce other parameters that affect the signal behaviour, these parameters can be easily incorporated in the proposed structure by using their respective probability density functions.
CHAPTER 4

Heterozygote detection

4.1 Introduction

Accurate identification of single nucleotide substitutions are crucial for diagnostic applications that require detection of mutated genes [32][33], for tissue compatibility testing prior to transplantation [34] and for analyzing forensic samples [36]. A commonly used method for detecting single nucleotide substitutions in DNA is fluorescence-based sequencing which was explained in Section 1.3. As described in Section 1.4.2, accurate detection of heterozygous sites or partial-mutations is limited due to the variability in fluorescence signals and overlapping of adjacent peaks of the same base type.

In Chapter 3, the ability of the hypothesis testing based L-DNA processor (derived in Section 2.3 and [56]) to detect homozygous sites was demonstrated. In this chapter, the L-DNA processor is used to estimate the local sequence in a complex comprised of homozygous and heterozygous sites. This method attempts to overcome limitations posed by both overlapping of adjacent peaks and inherent variability in the amplitudes of the homozygous and heterozygous peaks. Since the processor is derived based on a
formal statistical description of the data, it also provides the conditional probabilities
associated with each hypothesized sequence.

First, the need to associate relative probabilities (confidence) with the hypothe-
sized sequences are justified and then a relative probability measure is derived. The
use of the processor to identify heterozygous sites is illustrated using real DNA data.
This brings to light additional issues that need to be dealt with when the hypothesis
space is expanded to include heterozygotes and partial-mutations. The performance
of the processor as a heterozygote and mutation detector is assessed using a large set
of DNA data and the results are compared with that of a commercial base-calling
program (OpenGene by Visible Genetics Inc., Toronto, Canada). Finally, a general-
ized structure is proposed to overcome limitations associated with some of the model
parameters that are difficult to estimate.

4.2 Adaptation of the L-DNA processor

In order to detect heterozygotes, the hypothesis space of the L-DNA processor
is extended to account for the fact that in polymorphic sites, often two nucleotides
are present and hence two peaks may be present at the same base position in two
different lanes. If, for example, the hypothesized base type at a particular position is
M (a case where heterozygous members are A and C nucleotides – see Appendix B)
then the Θ function defined in (2.33) ensures that probabilities associated with half
height peaks in A and C lanes are computed by the likelihood function (2.34). Note
that the amplitude probabilities of heterozygous members are computed using the
expected level of heterozygous peaks in the respective lanes. Hence, in the case of
hypotheses with heterozygous sites, the joint likelihood of two peaks occurring at a
particular base position is evaluated. This differs from the likelihood evaluated at
homozygous sites, where the likelihood of a peak being present in only one of the four
lanes is evaluated. Using the statistics of the peak amplitudes and peak positions, one
can formulate the likelihood that the observed complex is generated by a particular
hypothesized sequence. Cost associated with each of the hypotheses is defined to
be the negative logarithm of the respective likelihood value. The hypothesis with
the lowest cost is chosen to be the sequence of nucleotides in the observed complex. The general steps involved in the identification of the local sequence comprising the complex is illustrated in Figure 4.1. In this document, the two bases which constitute a heterozygous site are referred as members of the heterozygous type.

Figure 4.1: Steps involved in detecting the local sequence in a complex using the L-DNA method. The particular hypothesized sequence with the highest likelihood (or the lowest cost) is selected by the processor. It should be noted that the set proportion of heterozygous peaks is hypothesis dependent.
4.3 Computation of confidence measure for each hypothesis

The size of the hypothesis space increases significantly when heterozygotes are included. This can lead to greater chance for confusion between competing hypotheses. Hence, there is a need to assess the relative confidence associated with each of the hypothesis. In this section, a method to compute relative probability of each of the hypothesis being correct is derived.

If the detection of a particular sequence $S$ is made based on a particular set of observations $l$, then the most likely sequence $S_i$ is obtained by maximizing the posterior probability:

$$S_i = \arg \max_{\{S_j\}} p(S_j|l) \quad j \in \{1, \ldots, i, \ldots, N\}$$

(4.1)

where $S_j$ is a particular hypothesized sequence. Using Baye’s theorem,

$$p(S_i|l) = \frac{p(l|S_i)p(S_i)}{p(l)}$$

(4.2)

Here, $p(l)$ is the prior probability of the observations and $p(S_i)$ is the prior probability of hypothesized sequence $S_i$. If the total number of possible hypotheses is $N$, then above equation can be written as:

$$p(S_i|l) = \frac{p(l|S_i)p(S_i)}{\sum_{j=1}^{N} p(l|S_j)p(S_j)}$$

(4.3)

It is clear that if all the hypothesized sequences $(S_j)$ are assumed to be equally likely, then:

$$p(S_i|l) = \frac{p(l|S_i)}{\sum_{j=1}^{N} p(l|S_j)}$$

(4.4)

Hence, the above expression provides the relative probability of hypothesis $S_i$ being correct when all the posterior probabilities $(p(l|S_j))$ are known.

Since the derivation of the L-DNA processor in Section 2.3 (Equation 2.34) pro-
vides the conditional probability of estimated signal amplitude for a particular sequence, the likelihood generated for a particular hypothesis has the form $p(l | S_i)$, where $l$ are the estimated signal and noise levels in the complex. Thus, the normalization suggested in (4.4) can be used to estimate the relative probabilities associated with each of the hypothesized local sequences in an observed complex. Similar method of normalizing likelihood measures is known in Artificial Neural Networks literature as soft max [57, 58].

4.4 Use of L-DNA processor as a proof reader for detecting heterozygous sites

The application of the L-DNA processor as a heterozygote detector is illustrated here through the use of two segments of DNA time trace (see Figures 4.2, and 4.3) obtained by sequencing a section of the HLA gene (HLA locus A exon 2) from genomic DNA. Thermo Sequenase$^TM$ and dye-labeled primer were used as part of the chemistry. The DNA time trace was obtained from Micro GeneBlaster$^TM$ sequencer (Visible Genetics Inc, Toronto, Canada). Here, the presence of a heterozygote was verified by sequencing the opposite strand of the DNA. The local bias level was first estimated and removed from the data. Isolated peaks which were identified with high confidence by a commercial base-calling algorithms (OpenGene$^TM$ by Visible Genetics Inc., Toronto) were used to estimate the following signal model parameters: mean and standard deviation of homozygous peak amplitude, mean and standard deviation of noise levels, peak shape parameters and peak width. These peaks were also used to estimate the mean peak separation and standard deviation of the peak-time jitter which are used in the local peak-time model [56].

Prior to computing the likelihood of the various hypotheses, the amplitude bias (known in communication literature as inter-symbol interference or ISI) from the known peaks in the vicinity of the complex were removed. For example, in Figure 4.2, the processor removes the amplitude bias of the A peak, which is located immediately to the left of the complex and extends well into the support region of the complex. Finally, the estimated values of the observed signal were used to evaluate
Figure 4.2: A segment of data with a heterozygous base type which is difficult to identify due to the presence of overlapping homozygous base of the same type. In this case the correct local sequence is GRC. The R base type is formed by the presence of 50% of G nucleotides and 50% of A nucleotides in the population DNA that was sequenced.

The likelihood associated with any local sequence that was hypothesized to occur in this complex. Since the proposed processor is a local processor, all possible hypotheses can be exhaustively tested. If only homozygous peaks are known to occur in a particular sequence and if the length of the segment under consideration is, say, 3 bases wide then there is a total of 64 possible hypotheses. Inclusion of heterozygous base types implies that each base position can be occupied by one of 10 possible types (4 homozygous bases and 6 heterozygotes). Hence, for a segment with 3 bases, there is a total of 1000 possible hypothesized sequences. This dramatic increase in the size of the hypothesis space can be significantly reduced by constraining the hypothesis space. If, for example, the hypotheses are further limited to contain at least one base from the lane in which the complex occurs and only one heterozygous base is
Figure 4.3: Second example used to illustrate the use of L-DNA processor to identify heterozygous sites. In this case the correct local sequence is ARC. The heavy tail of the complex implies that the peaks in this lane (A lane) are highly asymmetric.

assumed to occur in any one of the three positions in the complex, then the size of the hypothesis space\(^1\) reduces to only 169. Thus, \textit{a priori} information regarding the expected composition of a complex can be used to significantly limit the search space.

For the purpose of comparison, the same two segments were also processed using the Jansson-van Cittert iterative deconvolution (JID) algorithm described in Section 1.5.2 of Chapter 1. As suggested in Section 2.2, the relaxation factor was chosen such that convergence was guaranteed. Since the expected amplitude of peaks is unity, the parameter \(\gamma\) of the signal-dependent relaxation factor \(\beta\) (see Equations 1.12 and 1.16) could be set as high as 1. But to limit the amplification of noise peaks,

\(^1\)Total of 25 different purely homozygous triplets can be formed where 1, 2 or 3 bases are from the lane in which complex is present. For heterozygous peaks, constraining one member to be from the lane in which the complex occurs reduces the possible number of heterozygous base types to 3. Hence, 144 possible triplets exist where one of the triplet is a heterozygous base-type.
the relaxation factor was set to 0.4. The iterations were also limited to 50 to prevent amplification of noise peaks. It should be noted that the JID method is generally used only as a preprocessor to sharpen peaks and a heuristic method is employed to identify the number of constituent peaks. For this reason, only visual comparison is made between the results obtained using the JID algorithm and the L-DNA processor.

Table 4.1 lists the three hypotheses with the lowest costs (i.e. highest conditional probability) for each of the two data segments described above and the probabilities of these hypotheses being correct. For the first segment, the L-DNA processor correctly identified the local sequence as GRC\(^2\) with high probability (0.9999), while the competing hypothesis (GAC) has a much lower probability of being consistent with the observed data. The GAC hypothesis has a much lower probability of being correct for two reasons: (a) the presence of a half-height peak in the G lane, and (b) the signal level of the A peak being closer to the mean amplitude level of heterozygous peaks rather than homozygous peaks. The lane with the complex (G lane) and the estimated components of this complex at the expected base positions are illustrated in Figure 4.4. The result of processing the the same complex region using the Jansson-van Cittert algorithm is also shown in this figure. It is evident that the JID method results in two structures but the larger of the two components bares little resemblance to an isolated peak shape. This distortion occurs due to the inherent amplitude constraint which is incorporated in the JID algorithm.

For the second data segment (Figure 4.3), the L-DNA processor correctly identified the local sequence as ARC with a relative probability of 0.9932. Though, the top three hypotheses correctly identified the heterozygous peak (R) at the second base position, the competing hypotheses favoured the presence of a heterozygous peak in the first base position. In this the amplitude of the homozygous component at the first base position was much smaller than the expected value, which was almost unity. Hence, the second and third competing hypothesis favoured the presence of a heterozygous peak (small peak) at the first base position of the triplet. The lane with the complex and the estimated components of this complex at the expected base positions are illustrated in Figure 4.5. Unlike the results obtained for the first data segment, the

\(^2\)See Appendix B for naming convention.
amplitudes of the estimated homozygous and heterozygous components in the A lane do not significantly differ from each other. This further justifies why the competing hypotheses favour the presence of heterozygous types in the first base position. The performance of the JID method is very poor in this segment. The JID method generates several false peaks due to the presence of noise in the A lane. It should also be noted that the complex in Figure 4.3 (A lane) exhibited a pronounced downstream tail. This implied that the peaks in the A lane were highly asymmetrical. If the peak shape was assumed to be a simple Gaussian shape (as is commonly done in most base-calling programs [13, 11, 28]), this particular case would result in components with almost identical amplitudes (see Figure 4.6) and possibly cause their identification as two consecutive heterozygous sites.

### 4.5 Performance of the L-DNA Processor

The performance of the L-DNA processor is illustrated using two applications - detection of polymorphic sites in sequence traces of heterozygotes (for applications such as HLA typing) and early detection of mutations induced by pathogen attack (such as in HIV infection).
Figure 4.4: Results of processing data in Figure 4.2 using the L-DNA processor. The solid line indicates the observed G trace and the dashed lines are the two estimated components in the G lane that correspond to the most likely hypothesized sequence. The result obtained using a commonly used iterative deconvolution method (Jansson-van Cittert iterative deconvolution (JID) method) is also shown.

4.5.1 Detection of heterozygous sites

Heterozygous condition was artificially generated by mixing equal proportions of wild-type sample with a mutated variation where known base locations were substituted with known bases. Of these locations, two polymorphic sites were chosen where resolution was poor (see Figure 4.7). In these cases, accurate detection is particularly challenging because the smaller peak corresponding to a heterozygous member is present immediately adjacent to a homozygous peak of the same type and the two peaks overlap significantly (the complexes is about 40 bases away from the end of the read length). This overlap results in biasing of the amplitude and position of the smaller peak.
Figure 4.5: Results of processing data in Figure 4.3 using the L-DNA processor. The solid line indicates the observed complex in the A lane and the dashed lines are the two estimated components in the A lane that correspond to the most likely hypothesized sequence. Also shown is the result obtained using the Jansson-van Cittert iterative deconvolution algorithm (JID).

Five artificial mixture of samples were sequenced so that a total of 10 heterozygous segments were available for performance evaluation. The results of processing the time-traces using a commercial base calling program (OpenGene, Visible Genetics Inc, Toronto), which is specifically designed for use with the GeneBlaster sequencer, was also available for comparison purpose. The L-DNA processor was used to evaluate the likelihood associated with all possible triplet hypotheses for each of the selected complexes. Table 4.2 lists the best two hypothesized sequences for each of the ten cases and the associated probabilities.

Figure 4.8 illustrates the estimated components that constitute the observed complexes. Here the true sequences in the complexes are KTA and TYT, respectively.
### Table 4.2: Comparison of performance of the L-DNA processor with that of the commercial software for the 10 heterozygous cases. The top two hypotheses generated by the L-DNA processor for each data set and the relative probability of each of these hypothesis being correct is listed. The relative probabilities of the pairs of hypotheses (for each data set) do not add to unity since more than two hypotheses were considered for each data set.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>True Sequence</th>
<th>Commercial Software</th>
<th>Best L-DNA Hypotheses</th>
<th>Associated Probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>KTA</td>
<td>TGTA</td>
<td>KTA</td>
<td>0.9684</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>KYA</td>
<td>0.0193</td>
</tr>
<tr>
<td>2</td>
<td>KTA</td>
<td>KTA</td>
<td>KTA</td>
<td>0.9979</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>KWA</td>
<td>0.0013</td>
</tr>
<tr>
<td>3</td>
<td>KTA</td>
<td>KTA</td>
<td>KTA</td>
<td>0.9582</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>KYA</td>
<td>0.0222</td>
</tr>
<tr>
<td>4</td>
<td>KTA</td>
<td>KTA</td>
<td>KTA</td>
<td>0.9984</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>KYA</td>
<td>0.0009</td>
</tr>
<tr>
<td>5</td>
<td>KTA</td>
<td>KTA</td>
<td>KTA</td>
<td>0.9979</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>KYA</td>
<td>0.0012</td>
</tr>
<tr>
<td>6</td>
<td>TYT</td>
<td>TCT</td>
<td>TYT</td>
<td>0.9905</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TCT</td>
<td>0.0081</td>
</tr>
<tr>
<td>7</td>
<td>TYT</td>
<td>TCT</td>
<td>TYT</td>
<td>0.9981</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TCT</td>
<td>0.0015</td>
</tr>
<tr>
<td>8</td>
<td>TYT</td>
<td>TCT</td>
<td>TYT</td>
<td>0.9996</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TCT</td>
<td>0.0003</td>
</tr>
<tr>
<td>9</td>
<td>TYT</td>
<td>TCT</td>
<td>TYT</td>
<td>0.9991</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>WYT</td>
<td>0.0005</td>
</tr>
<tr>
<td>10</td>
<td>TYT</td>
<td>TCT</td>
<td>TYT</td>
<td>0.9463</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TCT</td>
<td>0.0451</td>
</tr>
<tr>
<td>Total Correct</td>
<td>4/10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>10/10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 4.6: Results of processing data in Figure 4.3 using the L-DNA processor assuming that the peak shape is symmetric. Notice that the resulting estimates are almost equal in amplitude though the first component is supposed to be a homozygous type and the second component is supposed to be a heterozygous type.

The commercial base-calling program correctly identified 4 out of the 5 KTA cases and one out of the 5 TYT cases. As evident in Table 4.2, the L-DNA processor correctly identifies all the cases. The L-DNA processor also provides an estimate of the underlying components of the complex. At first inspection, the identification of the KTA sequence is particularly challenging because the amplitude of the T member of the K base type is comparable to the amplitude of the homozygous T peak located immediately past the K base type. But the presence of a half-height G peak at approximately the same location favours the classification of the base type as K and not T. Notice that the half-height peaks in the G and T lanes do not occur exactly at the same location. The L-DNA processor accommodates for uncertainty in peak-times by estimating the amplitude of the components that could occur over a range of $\pm 2\sigma$ of
Figure 4.7: Illustration of two known polymorphic sites used to assess the performance of the LDNA processor. The complexes are located at approximate base position of 250 and the complete read length is approximately 300. (a) Here the heterozygous type is K. For a window of 3 nucleotides that include the complex, the correct local sequence is KTA. (b) Here the heterozygous type is Y. For a window of 3 nucleotides that include the complex, the correct local sequence is TYT.

the peak-time parameter. The dependence on a specific peak-time is reduced through the use of a marginalization operation [56].

In the TYT case, the commercial base-calling program failed to correctly identify the local sequence for two reasons. First, the amplitude of the T member of the Y base type was significantly smaller than the expected 50% amplitude. Hence, the C member of the Y type was easily mistaken to be a homozygous peak and the triplet was identified as TCT instead of TYT. The LDNA processor, on the other hand, considers the signal levels in all four lanes. The presence of a relatively smaller

\[3\text{Relative to homozygous C peaks}\]
C peak favours its detection as a member of a heterozygous base type. Second, the positions of the members of the Y type in the C and T lanes deviate significantly from each other. Hence, the C and the T members are identified by the commercial base-calling program as separate homozygous base types and this results in an insertion error (i.e. the wrongly identified sequence is often TCTT instead of TYT). Again, consideration of a range of peak-times by the L-DNA processor, allows it to easily cope with the variation in peak-times.
4.5.2 Detection of onset of mutations

Unlike heterozygous condition, where nucleotides at polymorphic sites are present in approximately equal proportions, the proportion of nucleotides present at a base location with substitution-type mutation may not be equal. The extent of a pathogen attack will determine the proportion of nucleotides at mutation sites. In terms of the DNA time trace, the former case results in half-height peaks at polymorphic sites, whereas, the latter case results in different peak-heights in different lanes at the same base position. Though the underlying biological phenomena are different in the case of heterozygous condition and substitution-type mutation, the two cases can be encompassed by the same signal model and processor presented in Section 2.3. As illustrated in Figure 4.1, a mean peak level is estimated prior to evaluating the likelihood of each hypothesis. In the case of heterozygous condition, the expected height of each of the heterozygous peak is set to be half that of the estimated homozygous peak. For substitution-type mutation, the mean levels are set to expected proportion of cell mutation. For example, to detect the condition where only 25% of the population of cells are mutated at a specific base location, the proportion of mean levels of heterozygous peaks at this location is set to 75:25.

To evaluate the performance of the L-DNA processor as a mutation detector, wild-type and mutant variant of the same fragment\(^4\) of DNA were mixed in 65:35 and 75:25 proportions. These were used to simulate the scenario where 35% and 25%, respectively, of the population of samples were mutated. Hence, the distribution of peak heights at polymorphic sites were expected to be proportional to the mixing ratios. The same two complexes that were considered in Section 4.5.1 were analyzed.

In order to detect the above mutations with the same detector, the proportion of mean levels of hypothesized heterozygous peaks were set to 70:30, which is in the middle of the 65:35 and the 75:25 mixing proportions. If, for example, the hypothesized local sequence is KTA, the processor will test for G:T ratios of 70:30 and 30:70 at the first base position of the local sequence. For the hypothesis where G:T ratio is 70:30, the heterozygous peak's amplitude\(^5\) in G lane is set to 0.7 and that of the

---

\(^4\)Fragment considered in Section 4.5.1 was used here.

\(^5\)Note that all the lanes were normalized such that the amplitude level of isolated homozygous
heterozygous peak in T lane is set to 0.3.

Table 4.3 lists the performance of the L-DNA processor for the scenario where 35% of the sample population is mutated and compares the performance with that of the commercial base-calling program. The L-DNA processor correctly identified all 5 KTA complexes and 3 out of the 5 TYT complexes. The commercial base-calling software correctly identified 4 of the KTA complexes, but identified only one of the TYT complexes. Though the L-DNA processor wrongly identified the local sequences in Data Sets 13 and 14, the relative probability of the decision being correct is significantly lower than what is observed for the other data sets in Table 4.3. For the same reason, the correctly identified sequence in Data Set 18 should be further analyzed to ascertain the correct local sequences. In this case, the peak position of heterozygous member in the T lane (corresponding to K type) deviates by approximately 41% of the mean base separation, whereas, the estimated standard deviation of peak positions for this data set is only 14% of the mean base separation. Hence, the KTA hypothesis is assigned a lower relative probability compared to KTA complexes in other data sets. As evident in Table 4.3, the commercial algorithm failed to recognize the local loss in alignment that occurs near the Y base-type in the TYT complex. Hence, the Y type was recognized as separate C and T types. Since the L-DNA processor considers a range of peak-times for each base position, it easily circumvents this limitation.

Table 4.4 shows the performance of the L-DNA processor as a mutation detector for the case where only 25% of the sample population is mutated. In this case, the proposed processor correctly identified 3 out of the 5 TYT complexes and 3 out of the 5 KTA complexes. Among the KTA complexes analyzed, two of the data sets (26 and 29) were identified as consisting of the homozygous sequence GTA. In these two data sets the amplitude of the T member of the K type was comparable to the noise level, whereas, the height of the G member was very close to the height of a homozygous peak (due to the 75:25 mixing proportion of wild and mutant types). Hence identification of the G member as a homozygous peak was favoured. Among the TYT complexes, two data sets (21 and 23) are wrongly identified as consisting peaks were close to unity.
Table 4.3: Illustration of performance of L-DNA processor as a mutation detector for data with mixing ratio of 65:35. The subscripts of the hypothesized local sequences indicate the assumed mixing ratio. The performance of the commercial base-calling program is also presented. *In DataSet 13, the relative probability of TYT_{70:30} is 0.0640 and in DataSet 14, the relative probability of TYT_{70:30} is 0.0112.
of two successive heterozygous sites. In Data Set 21, the position of the smaller C peak (heterozygous member of the Y type) was observed to deviate by 57% of the mean base separation. This favoured the hypotheses with a heterozygous C peak in the first position. Since a T peak was also present in the first position, the type in this position was identified as Y (C and T). Data Set 23 exhibited large deviations in the peak-times of the T and C peaks (members of the Y type). This resulted in wrong compensation for contribution from adjacent peaks. In particular, a significant component of the T peak at the first base position was removed. Hence, the first base position in the triplet was wrongly identified as consisting of a heterozygous type. The deterioration in the relative confidence with which the L-DNA processor makes its decision in the 75:25 data set (see the last column of Table 4.4) is associated with two factors: (a) decrease in signal-to-noise ratio (SNR) of the DNA peaks corresponding to the 25% mutation and (b) larger relative error between the set model parameters and the observed data. For example, in Table 4.4 the model assumes a 70:30 distribution of peak heights corresponding to wild and mutant types at polymorphic sites, whereas, the observed range was as high as 85:15. In this extreme case, the difference between the set proportion in the detector and the observed proportion of mutations is 100% (30/15). This mismatch will reduce the confidence associated with identifying specific hypotheses. To further demonstrate this point, when the KTA segments were analyzed assuming that the proportion of heterozygous peaks was 50:50, only 2 out of the 5 complexes with 65:35 mixing proportion and none of the complexes with 75:25 mixing proportions were correctly identified. Hence, it is vital to be able to set the proportion of peaks at polymorphic sites to reflect the expected proportion in the observed data.

The chart presented in Figure 4.9 summarizes the performance of the L-DNA processor and compares it with that of the commercial software provided by Visible Genetics Inc. for each of the mixing proportions mentioned above.

4.5.3 Generalized L-DNA processor

Like any statistical detector, the L-DNA processor computes probabilities associated with each of the hypotheses assuming that the parameter values of the signal
Table 4.4: Illustration of performance of L-DNA processor as a mutation detector for data with mixing ratio of 75:25. The performance of the commercial base-calling program is also presented.
Figure 4.9: Performance summary of the L-DNA processor. Pairs of base types in parenthesis indicate the two types of bases that together constitute a heterozygous type present at a given base position. The ratios in parenthesis refer to the proportions in which the two nucleotides were mixed to generate the heterozygous type.
model accurately describe the observed DNA time-trace. In the L-DNA processor, all the parameters, except the proportion of heterozygous peaks, are dynamically estimated from each observed data. Therefore, it can accommodate large variations in these parameters. In other words, the structure is data adaptive. The only parameter that cannot be estimated from each of the observed data is the expected proportion of heterozygous peaks. Hence, the detection performance deteriorates whenever the assumed expected values and the observed values deviate from each other. In order to accommodate large variations in ratio of peak heights at mutation sites the following modified structure was developed.

In situations where the processor has to identify a polymorphic site as either a heterozygous type or as a site undergoing mutation due to a pathogen attack, a combined structure can be used. The combined structure can compute in parallel the costs associated with several sets of hypotheses. Each set of hypotheses corresponds to presence of bases in specific proportions (for example, 50:50 and 70:30). The set containing the hypothesis with the lowest cost is selected as the hypothesis space over which the decision is made. The hypothesis with the lowest cost implies that the proportion of peaks at the mutation site in the observed data is closest to the proportion assumed by the processor. This structure is referred to as generalized L-DNA (see Figure 4.10). The advantage of selecting one of the sets of hypotheses over pooling hypotheses from different sets is that the size of the hypothesis space can be limited to the size of each set. Hence ambiguities, which are introduced by large hypothesis space, can be avoided.

Performance of the generalized L-DNA processor was evaluated using all of the KTA complexes considered in previous sections (5 segments of KTA complexes in 50:50, 65:35 and 75:25 proportions each). For these data, two sets of hypotheses were considered, namely, 50:50 mixture of wild and mutant types and 70:30 mixture of wild and mutant types. For the 50:50 data, the 50:50 hypothesis set contained the hypothesis with the lowest cost. Similarly, for the 65:35 and 75:25 data sets, the 70:30 hypothesis set contained the hypothesis with the lowest cost. Hence, the overall performance of the generalized L-DNA processor was identical to the performance of the L-DNA processor illustrated in Tables 4.2, 4.3, and 4.4. In other words, all the
Chapter 4 \ Heterozygote detection

Figure 4.10: Generalized L-DNA processor: Implementation of a combined structure to detect not only possible polymorphic sites but also classify them as either pure heterozygotes or sites exhibiting early stages of mutation in various proportions.
KTA complexes in 50:50 and 65:35 data sets were correctly identified, while only 3 out of the 5 KTA complexes were correctly identified in the 75:25 data sets.

The advantage of the generalized L-DNA processor is that each sub-processor can be optimized for the detection of a specific ratio of bases at polymorphic sites. To distinguish between pure heterozygotes and polymorphic sites with unequal proportion of bases (indicative of onset of mutations) it is sufficient to use just two sub-processors — one optimized for detecting pure heterozygotes and another optimized for detecting onset of mutations. The heterozygote detector would assume a 50:50 distribution of heterozygous peaks and the mutation detector would assume a 70:30 distribution of peaks at mutation sites.

4.6 Discussion

In Chapter 3, the error rate of the L-DNA processor was assessed using simulated data. This method requires generation of several realizations of the DNA time series to assess the accuracy with which the L-DNA processor identifies the local sequences. In this chapter, a relative probability measure is described that is based on the observed DNA time trace and does not require simulation. The fundamental assumption of both these methods is that the signal model accurately describes the observed DNA time-trace. Neither of these methods account for the scenario where the observed DNA time-trace departs significantly from the assumed model.

The extension of the L-DNA processor to detect heterozygous sites illustrates how the processor can be easily adapted to detect specific cases. This is achieved by simply extending the hypothesis space of the processor. Another such application of the L-DNA processor would be the detection of a common chemical artifact in DNA sequencing known as hard stops [41]. In this condition, fragments with the same length are generated in all four lanes. Hence, peaks appear in all four lanes at the same base position. This can be easily detected using the L-DNA processor by including a case where peaks are hypothesized to occur in all four lanes a particular peak position. Hence, the L-DNA processor can be used to remove artifacts prior to base-calling.
In comparing the performance of the L-DNA processor with the commercial base-calling program, the parameters of the commercial program were assumed to be set at default values. It might be possible to improve the performance of the commercial base-calling program by fine tuning the parameter values used by the program. Yet, it is anticipated that the difference would not be significant to affect the general conclusions drawn in this chapter.

Finally, the generalized structure presented in Section 4.5.3 is not without limitations. For example, the number of processing stages (i.e. number of distinct proportions of wild and mutant types) cannot be arbitrarily chosen since the variability in peak heights limit the accuracy with which mutation sites can be further classified based on the proportion of peak heights at these sites. But the structure with only two processors is shown to be sufficient to distinguish between pure heterozygous sites and sites with partial-mutations.
5.1 Thesis Summary

A processor that is suited for application as a local post-processor for existing DNA base-calling programs has been developed. The processor is capable of identifying the local sequence in regions that have been tagged as ambiguous by conventional base-calling programs. The ability of the proposed processor to resolve overlapping peaks has been studied using both simulated data and real DNA data. The robustness of the processor in correctly identifying peaks in low resolution regions of DNA data was illustrated through its use as a sequential decoder to extend the length of DNA that could be read in a single sequencing run. The proposed processor was also extended to detect heterozygotes and sites with partial mutations. The performance of the processor in detecting heterozygotes and mutations was illustrated through the use of a large set of real DNA data. In regions where peaks overlap, the performance of the proposed structure is shown to be significantly better than a commercial base-calling program.
5.2 Thesis contributions

- *Development of the first local statistical processor for DNA sequencing*: This work has resulted in the development of a practical post-processor that uses statistical information inferred from high resolution data to detect the local sequence in regions with low resolution. The statistical DNA model has been extended to represent data exhibiting two clinically relevant conditions: heterozygous sites and partial-mutations. Flexible nature of the proposed structure allows it to cope with uncertainty in model parameters. For example, a generalized structure has been developed using the L-DNA processor to distinguish between heterozygous sites and sites undergoing partial-mutation. The hypothesis-testing approach facilitates easy incorporation of special cases (such as heterozygous sites and hard stops) into the suite of conditions that can be detected by the processor.

- *Use of conditional probabilities to reduce the dimensionality of the detection problem in DNA data*: Marginalization of conditional probabilities provides a means of reducing the influence of those conditional parameters that cannot be easily estimated. This technique has been successfully used to reduce the number of signal model parameters that need to be estimated for accurate detection of local DNA sequences.

- *A unified estimation-detection technique for accurate detection of overlapping peaks*: Base-calling programs published to-date (with the exception of [48]) rely on two distinct stages for identifying peaks in low resolution regions: (i) deconvolve peaks in overlapping regions, and (ii) identify the sharpened peaks using heuristic methods that are not optimal. The proposed method estimates the amplitudes of the constituent peaks and uses these values to associate a likelihood with each of the hypothesis. The amplitude estimation process is hypothesis dependent. Hence, the hitherto distinct processing stages are unified in the L-DNA processor.

- *Provide relative probabilities (confidence) for the processor's decision*: The like-
likelihood values associated with each of the hypothesis can be used to calculate the corresponding relative probability provided all possible likelihood values are known in advance. Since the L-DNA processor is used as a local processor, all possible hypotheses can be exhaustively evaluated without any significant computational burden. Clinically this has two advantages: (i) the relative probabilities will be invaluable to the human expert who would need to draw higher level conclusions from the identified sequence, and (ii) the confidence values also provide the ability to assess the need for human intervention in editing the results of automated DNA sequencing.

- **Industrial and clinical significance:** The performance of the L-DNA processor has been shown to be significantly better than commonly used algorithms and commercial DNA base-calling programs. This significant improvement in performance of the processor has warranted commercial exploration of this technology. The author of this work will be further assessing the feasibility of integrating this technology with a commercial DNA sequencing system.

### 5.3 Avenues for future research

This thesis provides a general structure for a local DNA processor. The processor uses *a priori* information that is gathered from regions in DNA data with high resolution to identify sequences in low resolution regions. In order to simplify the structure of the processor, several simplifying assumptions were made with regard to the signal model. For example, the difference between successive positions of the peaks were assumed to be identically distributed independent random variables. Yet, previous work have shown that these differences are highly correlated. If the difference in peaks times are assumed to be highly correlated, then the limits of the multiple integrals in the likelihood function (2.34) would reduce dramatically. The significance of this extra information on the simplification of the likelihood function should be evaluated further using real DNA data.

The performance of the L-DNA processor is highly dependent on the accuracy with which the various statistical parameters are estimated. In this work standard
methods were used to evaluate expected values and standard deviations of the model parameters. Methods used to estimate statistical parameters of small sample sizes could be explored to improve the overall accuracy of the processor.

Another means of improving the accuracy and consistency of estimated model parameters is to improve the preprocessing steps. The development of the processor, for example, assumed that the four lanes were optimally aligned with respect to each other by a base-calling program. Similarly, base-line subtraction has a direct impact on the estimated peak amplitudes, peak widths and noise levels. Hence, explorations could be pursued to improve these preprocessing steps.

From a practical implementation point of view, the size of the hypothesis space of the processor can be significantly reduced in regions where all the hypotheses are not equally likely. Higher level knowledge about a specific genetic condition can be used to assign different weights to each of the hypothesis and completely ignore some other hypotheses. These modifications could be implemented when accurately screening for specific diseases where the various genetic variations are known a priori. Hence, in the post human genome era, where various diseases are being characterized at the genetic level, this can prove to be a valuable proof reading tool.
Bibliography


APPENDIX A

Implementation and analysis of JID algorithm

A.1 Implementation details

The Jansson-van Cittert algorithm, which uses a signal-dependent relaxation factor, and the general van Cittert algorithm, which uses a constant relaxation factor, were implemented using the matrix form described in (2.14). The area of the PSF was normalized to one so that energy of the estimated signal \( g \) was preserved in successive iterations. The point spread function \( g \) was shifted such that it introduced no time-shift when convolved with the estimated signal (i.e. \( g \) was shifted so that its peak was at the origin). Since the \( D \) matrix represents the linear convolution by the PSF, it becomes a block-Toeplitz matrix [59]. Both the observed signal \( y \) and the point spread function \( g \) were zero padded so as to avoid edge-effects during convolution operation. This also resulted in \( D \) matrix being block-circulant. The characteristic of the block-circulant matrix is that the result of linear and circular convolution are the same [9]. Finally, the number of iterations was selected manually based on visual
inspection of the estimated signals. The iterations were generally terminated when spurious peaks were excessively amplified.

A.2 Empirical analysis of JID algorithm

The signal dependence of the relaxation factor $\beta$ does not lend itself favourably for either eigenvalue analysis or frequency domain analysis. For this reason, the convergence behaviour of the Jansson-van Cittert method was analyzed empirically using simulated data.

To illustrate the successive peak sharpening property of the JID algorithm, a complex was generated by convolving two impulses located 10 samples apart with a Gaussian shaped PSF. The full width at half maximum (FWHM) value of the PSF was 10 samples. Figure A.1 illustrates the complex and the sharpened peaks at several iterations. It is evident that in the limit, the algorithm converges towards two impulses. This was in agreement with the assumption of the algorithm – the observed signal is the convolution of a point spread function and an impulse train representing the location of the peaks.

Figure A.2 compares the performance of the general van Cittert algorithm with that of the Jansson-van Cittert algorithm. The fixed relaxation factor for the general method was selected so that the condition in (2.21) was satisfied. For the PSF, the largest eigen value was 1 and hence $\beta$ was set to 1.75 (which satisfies (2.21)). This figure shows that it is possible to set the constant relaxation factor $\beta$ to obtain the same performance as Jansson's modified algorithm. The obvious advantage of the JID algorithm is that negative excursions of the signals, which are not realistic for DNA data, are suppressed. This implicit positivity constraint is included in the signal-dependent relaxation factor given in (1.16), which is repeated here:

$$\beta_i = 1 - \frac{2}{\gamma} \left| x_i^{(k)} - \frac{\gamma}{2} \right|$$  \hspace{1cm} (A.1)

It is evident from above expression that for negative estimates of the original signal $x$ and for cases where amplitude of $x$ exceeds $\gamma$, the relaxation factor takes on negative
Figure A.1: Illustration of peak sharpening property of JID algorithm.
values. Hence, an opposite correction is applied to prevent signals from taking on any values outside of the 0 to $\gamma$ range. In this study, $\beta$ was set to 2 so as to prevent distortion of the estimated signal.

The extent to which the JID method could resolve overlapping peaks was explored by synthesizing several complexes with varying amounts of overlap. Figure A.3 illustrates the results of this exploration. Here, peak separation between adjacent peaks was reduced from 10 to 7.5 and then to 5, while the FWHM value was set at 10 for all the cases. All three complexes were processed using 50 iterations of the JID algorithm. It is evident that the JID method cannot resolve peaks that are separated by less than 50% of their peak widths. This is a serious limitation in the case of DNA data where such loss in resolution is typically observed to occur towards the end of a sequencing run.
Figure A.3: Performance of JID algorithm with respect to resolution of signals. Here, resolution $R$ is defined as the ratio of peak width and peak separation.
APPENDIX B

Nomenclature used to represent DNA sequence

Following table is part of the nomenclature suggested as a standard in [60]. The table lists the nucleotides present and the naming convention used to identify them when representing DNA sequences.

<table>
<thead>
<tr>
<th>Base type</th>
<th>Identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adenine</td>
<td>A</td>
</tr>
<tr>
<td>Cytosine</td>
<td>C</td>
</tr>
<tr>
<td>Guanine</td>
<td>G</td>
</tr>
<tr>
<td>Thymine</td>
<td>T</td>
</tr>
<tr>
<td>A and C</td>
<td>M</td>
</tr>
<tr>
<td>A and G</td>
<td>R</td>
</tr>
<tr>
<td>A and T</td>
<td>W</td>
</tr>
<tr>
<td>C and G</td>
<td>S</td>
</tr>
<tr>
<td>C and T</td>
<td>Y</td>
</tr>
<tr>
<td>G and T</td>
<td>K</td>
</tr>
</tbody>
</table>

Table B.1: Nomenclature suggested by International Union of Biochemistry.