DNA BASE-CALLING TECHNIQUES

Fadi Odeh, Candidate for the Master of Science Degree

American University of Sharjah, 2008

ABSTRACT

The availability of substantial amounts of DNA sequence information has begun to revolutionize the practice of biology. So it is obvious that manual sequencing output is not adequate to keep pace with the growing demand and is far from what is required to obtain the 3-billion-base human genome sequence.

To avoid this difficulty, replacing manual sequencing with an automated one is essential, and it is particularly important that human involvement in data processing be significantly reduced or eliminated. Progress in this respect requires both improving the amount of error-free data being processed, as well as the reliable accuracy measures to reduce the need for human involvement in error correction. Here, we precede one step toward that goal: a basecalling program for automated DNA sequencing, with improved accuracy.

The major goal of this thesis is to develop a new basecalling technique to improve the efficiency of the DNA sequencing process. Improved efficiency will be achieved by increasing the average length of error-free sequences and enhancing the base identification

process at the beginning and end of the DNA sequences. This will greatly increase sequencing throughput and reduce both cost and error associated with the current DNA sequencing process. ABI machines (Applied Bio-systems Incorporated sequencing machines) are currently the major source of reading DNA data. They are capable of producing sequences of 1000 bases in length (bases produced by PCR (Polymerase chain reaction)). These machines are associated with basecalling software, the most advanced software is called KB Basecaller v1.4 and it is publicly used by the sequencing community because of its reliability and accuracy. It can produce impressive results of 500~600 error-free sequences. The error-free sequences are normally located in the middle of the 1000 base length where the data is clear, and bases are easily distinguishable. However, the bases at the beginning and end of a 1000 base sequence are obscure and difficult to identify. The base calling error in these regions is relatively high. Thus the average basecalling error over a 1000 base sequence is between 3.5 and 6%. The foundation of this proposed research is based on a new base-calling program related to combining signal processing and pattern recognition systems which includes the following steps: noise filtration, baseline adjustment, mobility shift correction, feature extraction and the development of an intelligent basecalling algorithm. The new algorithm will be tested and validated on a number of pre-sequenced DNA sequences.

Combining Gaussian Mixture Models and Hidden Markov Models (GMM-HMM) classifier will be used as a classification model for the recognition of the DNA bases based on its several advantages over other classifiers in that they do not require heavy training, they are very simple to implement with the number of classes, and they ensure the coverage of the statistical properties of the data using Gaussian distribution.

DNA sequence information is critical to understand genetic variations that can influence both disease, and genetic interactions, which in turn can influence drug efficacy. As such, automated sequencers play a vital role in the drug discovery process.

CONTENTS

ABSTRACT i

LIST OF FIGURES v

CHAPTER 1 DNA SEQUENCING BACKGROUND 1

1. BIOINFORMATICS 1
   The Goals of Bioinformatics 1

2. INTRODUCTION TO DNA 2

3. HOW GENES CAUSE CANCER 3
   Oncogenes 4
   Tumor Suppressor Genes 4
   DNA Repair Genes 5

4. SUMMARY 6

CHAPTER 2 DNA SEQUENCING PROCESS 7

1. DNA SAMPLES PREPARATION 7

2. ELECTROPHORESIS 8

3. Data Flow 10
   Data Frame 10

4. SIGNAL PROCESSING 11
   Signal Pre-Processing 12
   Errors in Signal after Pre-Processing 15

5. BASECALLING 17

6. SUMMARY 18

CHAPTER 3 PATTERN RECOGNITION SYSTEMS 19

1. INTRODUCTION 19

2. DESIGN OF PATTERN RECOGNITION SYSTEMS 19

3. SELECTION OF MODEL PARAMETERS 20
   Gradient Descent 20
   Expectation-Maximization 22

4. PATTERN RECOGNITION MODELS.. 26

Gaussian Mixture Models.. 26

Artificial Neural Networks.. 30

Neuron Model... 31

Network Structure... 33

Error Backpropagation Network....................... 34

Polynomial Classifiers.. 37

Hidden Markov Models (HMMs)...................................... 39

Introduction... 39

Markov Chains... 39

HMM Elements... 41

HMM Model... 42

5. HYBRID PATTERN RECOGNITION MODELS............................ 48

Combining GMMs and HMMs... 49

Combining ANNs and HMMs... 51

Combining Polynomial Classifiers and HMMs....................... 52

6. SUMMARY... 56

CHAPTER 4 COMBINING SIGNAL PROCESSING AND PATTERN RECOGNITION MODELS TO SOLVE DNA SEQUENCING PROBLEM 57

1. INTRODUCTION.. 57

2. SIGNAL PROCESSING.. 58

Noise Filtration... 59

Discrete Wavelet Packet 60

Baseline Correction.. 61

Mobility Shift Correction 62

3. FEATURES SELECTION.. 65

4. HMM TOPOLOGY 66

5. RESULTS EVALUATION AND CONCLUTION 68

Evaluation Results... 69

Conclusion 74

6. SUMMARY... 75

REFERENCE LIST.. 76

LIST OF FIGURES

Figure 1.1: From genes to proteins ……………...………………………...... 2

Figure 1.2: Gene coding ……………………………………...………….. 3

Figure 1.3: Germline mutation ……………………………....………… 4

Figure 1.4: Acquired bone cells …………………………………………….. 4

Figure 1.5: Gene mutation …………………………………………...…... 4

Figure 1.6: Mutation in HBB gene …………………………………….…. 5

Figure 1.7: Difference between normal blood cells and sickle blood cells …. 5

Figure 2.1: Cyclic sequencing ………………………………………………. 7

Figure 2.2: Chain termination DNA sequencing …….………………….……... 8

Figure 2.3: Electrophoresis process representation …….……………………... 9

Figure 2.4: Normalized fluorescence emission spectra for each of the four dyes following excitation at 488nm……...…………………….. 10

Figure 2.5: Charge-coupled device camera organization ……………….…. 11

Figure 2.6: Fluorescence pattern …………………………………………… 11

Figure 2.7: Background noise affects the basecaller…...…………………….. 12

Figure 2.8: The problem of spectral cross-talk between DNA sequence data channels…………………...……………………………….... 13

Figure 2.9: Cross-talk problem removal ……….…..………………………. 13

Figure 2.10: The problem of baseline offset between DNA signals ………..... 14

Figure 2.11: The effect of mobility shift problem in DNA signals ……..…… 14

Figure 2.12: Mobility shift correction …….……….………………………… 15

Figure 2.13: Extra space between A and G was ignored by basecaller……… 15

Figure 2.14: Extra space between A and G was picked as error by basecaller.. 16

Figure 2.15: Extra space between A and G was picked as a real base by basecaller ………………………………...………………… 16

Figure 2.16: Single nucleotide polymorphism SNP ...……………...……... 16

Figure 2.17: Two peaks above each other ……………………………….. 17

Figure 2.18: Dye blob problem……………………………………………... 17

Figure 3.1: Pattern recognition cycle……………………………………... 19

Figure 3.2: Gradient descent ……………………………...……….…….. 21

Figure 3.3: Weights update using gradient descent………………………….. 22
Figure 3.4: Expectation-Maximization overview…….……………………… 25
Figure 3.5: Multiple gaussian components when modeling a distribution…… 26
Figure 3.6: Typical multilayer perceptron model……………………….…… 30
Figure 3.7: Neuron model……………………..……………………....…… 31
Figure 3.8: Unit step transfer function……………………………………… 31
Figure 3.9: Linear transfer function………………………………………… 32
Figure 3.10: Sigmoid transfer function……………………………………… 32
Figure 3.11: Softmax transfer function…………………………………….….. 32
Figure 4.1: Raw data signal generated from the sequencing machine…….….. 58
Figure 4.2: Discrete wavelet transformation of 3 levels …………………….. 60
Figure 4.3: Raw data signal after noise filtration and baseline correction …… 62
Figure 4.4: Raw data peaks before shifting …………………………………. 64
Figure 4.5: Raw data peaks after shifting ...…………………………….…. 64
Figure 4.6: DNA bases features. The height and width extracted for each peak……………………………………………………….…...….. 65
Figure 4.7: Mutated raw data peak …………………………………………. 66
Figure 4.8: HMM topology to model DNA bases transitions ………………. 66
Figure 4.9: Markov model representing states transition of base peak in time: rising, width, and top ………………………………………. 67
Figure 4.10: Final HMM structure that will be used for recognition ………... 67
Figure 4.11: Simplified HMM model ……………………………………..…. 68
Figure 4.12: The results of comparing ABI software with GMM-HMM basecaller …………………………………………………... 71
Figure 4.13: The results of comparing PHRED with GMM-HMM basecaller. 72