A DIGITAL SIGNAL PROCESSING APPROACH TO ANALYZE GEL ELECTROPHORESIS IMAGE

NOOR EZAN BINTI ABDULLAH

A project report submitted in partial fulfilment of the
requirements for the award of the degree of
Master of Engineering (Electrical - Mechatronics and Automatic Control)

Faculty of Electrical Engineering
Universiti Teknologi Malaysia

NOVEMBER 2009

Dedicated with deepest love to:

My beloved family for their support, guidance and love.

My dearest friends for being there whenever I needed them.

ACKNOWLEDGEMENT

In the name of ALLAH, the Most Beneficent, the Most Gracious and the Most Merciful who has given me patience in completing this project.

Firstly, I would like to express my appreciation to my supervisor Dr Zuwairie Ibrahim for his contribution and encouragement upon accomplishment of this project. My gratitude also goes to Mr Amir, Tutor UTM , Mr Rozaimi and Mr Ahmad NurAlifah for their willingness in sharing their knowledge that has made this work excitably fruitful.

Also not to forget my family, all my friends and for those who have helped me either directly or indirectly for their guidance and willingness in sharing knowledge towards the completion of this project. All of them had given useful ideas towards the accomplishment of this project and always spend their time especially when I face problems. I am greatly indebted for their help and willingness in reading and correcting my report.

Last but not least, I wish to express my gratitude to the mighty God for giving me good health in completing this project and throughout my studies. I am sure and believe that without the contributions from those mentioned and also without good health, it is hard and impossible for me to fulfill the objective of this project.

ABSTRACT

Gel Electrophoresis (GE) is a widely used technique to separate DNA according to their size and weight, generates images that can be analyzed automatically. The separated DNA fragments or proteins of different molecular weights will give a series of bands with positions corresponding to the molecular weight. Image analysis of the gels removes much of the subjectivity of manual comparison of band position and intensity between samples. Briefly, this project presents a semiautomatic image processing techniques attempts to detect lanes, bands and length or size estimation of the bands. In this project, the routines are fully written in MATLAB R2008a. This project consists of three stages. The first stage is pre-processing, which involves conversion of RGB image into greyscale image. The second stage is to identify the number of lane in the image including the lane for marker and also to distinguish between the lane of marker and unknown DNA samples. The final stage is the identification of the length of DNA molecules each band in the lane based on data provided by the DNA marker.

ABSTRAK

Elektroforesis gel adalah salah satu teknik yang sering digunakan untuk mengasingkan DNA berdasarkan saiz dan berat di mana imej yang dihasilkan boleh dianalisa secara automatik. Pengasingan bahagian pecahan DNA atau protein daripada berat molekul berlainan akan memberi satu siri daripada belang-belang dengan kedudukan berdasarkan berat molekul ini. Analisis imej untuk gel – gel ini banyak menyingkirkan manual subjektiviti untuk perbandingan daripada kedudukan belang dan kesungguhan di antara sampel. Secara ringkasnya, projek ini menunjukkan pemprosesan imej secara semi automatik yang cuba untuk mengesan lorong-lorong , belang-belang dan penaksiran panjang atau saiz belang – belang ini. Di dalam projek ini, rutin sepenuhnya ditulis di dalam MATLAB versi R2008a. Projek ini terdiri daripada tiga peringkat. Peringkat pertama adalah pra pemprosesan di mana melibatkan penukaran dari imej berwarna (RGB) kepada imej berwarna kelabu. Peringkat kedua adalah untuk mengenal pasti bilangan lorong yang terkandung di dalam imej termasuk lorong dari penanda dan juga untuk pembezaan di antara lorong dari penanda dan sample-sampel DNA yang tidak diketahui. Peringkat yang terakhir adalah pengesanan panjang atau saiz DNA molekul dari setiap belang berdasarkan data yang dibekalkan dengan syarat daripada penanda DNA.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

A	-	Adenine
bp	-	bits per pixel
bpp	-	base pair
BMP	-	Bitmap image
C	-	Cytosine
DNA	-	Deoxyribonucleic acid
EM	-	Expectation – Maximization
FORTRAN	-	The IBM Mathematical Formula Translating System
G	-	Guanine
GE	-	Gel Electrophoresis
GUI	-	Graphical User Interface
IMTOOL	-	Image Processing Toolbox
JIT	-	Just –In-Time
JPEG	-	Joint Photographic Experts Group
KB	-	kilobyte
MATLAB	-	MATrix LABoratory
MB	-	Megabyte
MW	-	Molecular Weight
OOP	-	Object – oriented Programming
PFGE	-	PulsedField Gel Electrophoresis
RGB	-	Red, Green, Blue
RNA	-	Ribonucleic acid
ROI	-	Region of Interest
Std	-	standard deviation
T	-	Thymine
TLC	-	Thin Layer Chromatography
UV	-	Ultraviolet

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Project Background

Electrophoresis is an electrochemical separation process in which biological molecules, such as protein or RNA/DNA fragments, are made to migrate through a specific substrate, usually a polyacrylamide gel, under the influence of an electric current. The technique can be used to separate mixtures of molecules on the basis of their molecular size, by making use of their electric charge differences. This difference, under the electric field charge, causes individual biological materials of the same size to migrate to discrete positions within the bed of polyacrylamide medium. Collection of these multiple positions in a linear fashion presents the separation of mixed biological materials into specific electrophoresis profiles. It has wide application in DNA sequencing, and in studying variation in the qualitative and quantitative separation of proteins or nucleic acids obtained from different sources. Scientists use electrophoresis to derive information about the substances under study, such as comparing the composition of samples, or quantifying the amount and properties of the different constituents present in a collection of samples. Electrophoresis has many variants, including one or two-dimensional electrophoresis, electrofocusing, isotachophoresis, and several forms of immunoelectrophoresis [1].

Agarose Gel Electrophoresis is a commonly used method of separating molecules of based on their charge, size and shape. It is especially useful in separating DNA and RNA fragments are made to migrate through a specific substrate as illustrated in Figure 1.1 [2].

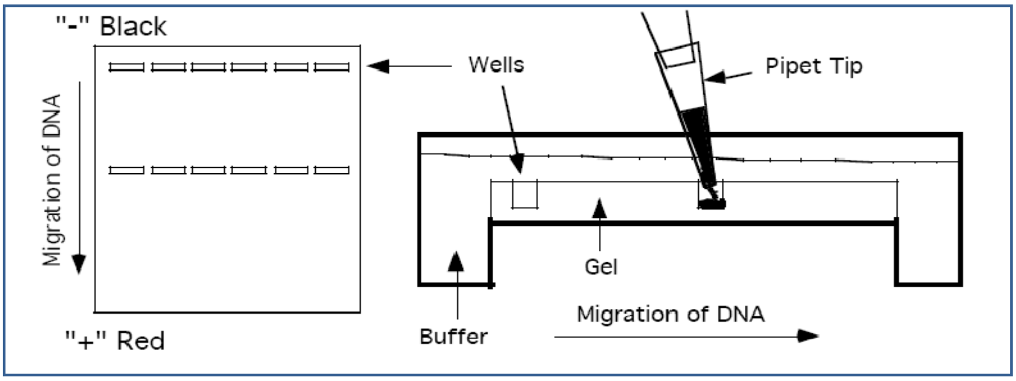

Figure 1.1: The migration process of DNA.

Gel Electrophoresis (GE) is an important tool in genomic analysis. GE result can be presented using images. Figure 1.2 shows the original image of GE where the image contains several vertical lanes, each lane corresponding to one sample. Each lane contains some horizontal bands and each band represents a part of the sample [3, 4].

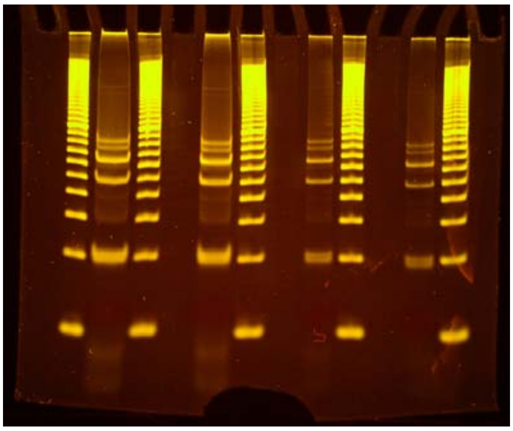

Figure 1.2: Original image of Gel Electrophoresis

Present works of identifying, processing and analysis of GE images may consume time and costly which is needed a sophisticated equipment for more details parameters. With the advancement of computer technology, processing and analysis of any gel electrophoresis images can be visualized and can be very cost effective [5] as well as it is the quickest way to obtain quantitative data from gels [5, 6]. In this study, image processing technique is applied to identify between the lane marker and unknown DNA samples and finally estimated the length of the DNA molecules each band.

1.2 Problem Statement

Many factors that could affect the image quality, such as applied voltage and field strength, pulse time; reorientation angle, agarose type, concentration, the buffer chamber temperature and others related effect [3, 4]. Furthermore, the locations of the lanes and the size of the lanes in the image are different. All these factors make the lanes extraction and comparison difficult.

Besides that, comparing two lanes in a gel electrophoresis image is usually a complex process as the subjectiveness of human visual perception and the factors related to the experiments may lead to different conclusions, even if the same material is applied. An automatic analysis of the band pattern of a lane could enable the evaluation of many parameters that are usually ignored by human analysis. However, basic tasks such as the identification of lanes in a gel image, easily done by human experts, emerge as problems that may be difficult to automate [7].

Thus by using the image analysis techniques, hopefully this application provides a relatively quick and inexpensive method for biologist in order to detect lane as well as estimated the length of DNA band.

1.3 Objectives Of Project

The objectives are:

a) To identify the lane in the gel electrophoresis images.

b) To distinguish between the lane marker and unknown DNA samples.

c) To develop software that is capable to calculate or estimate the length of DNA molecules each band in the lane.

1.4 Scopes Of Project

There are several scopes that should be covered to determine the boundary of this project. The scopes are:

a) To study gel electrophoresis images in order to identify the lane and the marker.

b) The image processing part for this study is done by using Image Processing Toolbox in MATLAB R2008a Software.

c) To analyse the electrophoresis image as shown in Figure 1.3

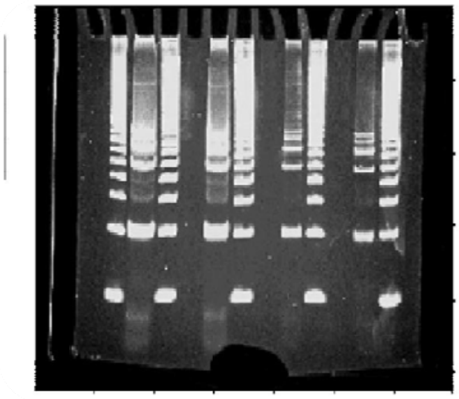

Figure 1.3: Gel electrophoresis image.

1.5 Significant Of Project

The significant of this project is to help the users (researchers and biologists) to observe, analyse and later convey his or her thoughts about detecting lanes as well as to estimate its length and give conclusion in a quick and intuitive way.

1.6 Thesis Organization

This thesis is organized into five chapters.

Chapter 1 will present the introduction of the project, which is brief information and scope of the project is discussed. Several facts about the previous work by other researchers also been touched.

Chapter 2 contains literature review and detail about the information and scope of the project. It is also briefed some of the MATLAB application that involved in the gel electrophoresis images analysis.

Chapter 3 discusses briefly on the methodology and results for the project. This chapter reveals about some theories and the algorithm procedures. While, for the results, all graphs, tables and comments were included.

Chapter 4 will present the discussions of the project. This chapter discussed about the MATLAB software and the algorithm used.

Chapter 5 is for conclusion on the study and some review for the future recommendation works will be explained.

1.7 Gantt Chart

The Gantt chart is a tool for planning and scheduling that used in this project. It enables an initially scheduling project activity and then monitors progress over time by comparing planned progress to actual progress. In this project, there are two Gantt chart are developed; Gantt chart for Project Proposal and Final Project respectively as shown in tables below.

Table 1.1: Project Proposal schedule

		PROJECT PROPOSAL (WEEKS)															
NO.	TASK/ACTIVITIES	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	Discuss proposal	▨															
2	Project literature/Background Research		▨	▨	▨												
3	Data collection for project analysis				▨	▨											
4	Learning and getting familiar with MATLAB					▨	▨	▨									
5	Submit project synopsis							▨									
6	Apply image processing techniques							▨	▨	▨	▨	▨					
7	Submit presentation material and preparation for presentation											▨	▨				
8	Presentation													▨			
9	Final report writing													▨	▨	▨	▨
10	Submission final report																▨

Table 1.2: Final Project schedule.

		FINAL PROJECT (WEEKS)															
NO.	TASK/ACTIVITIES	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	Design and develop algorithm for lane detection and length estimation	■	■	■	■	■	■	■									
2	Implement the algorithm and validate data			■	■	■	■	■									
3	Final validation data								■	■	■						
4	Preparation for presentation										■	■					
5	Presentation												■				
6	Thesis writing													■	■	■	
7	Submission thesis																■