DNA Sequence Reading by Image Processing

A thesis presented in partial fulfilment of the requirements for the degree of

Master of Science

in Computer Science at

Massey University

Supervised by

Dr Donald Bailey

and

Dr John Hudson

Fan Baozhen

1993

To my mother, father
and sisters

Abstract

The research described in this thesis is the development of the DNA sequence reading system.

Macromolecular sequences of DNA are the encoded form of the genetic information of all living organisms. DNA sequencing has therefore played a significant role in the elucidation of biological systems. DNA sequence reading is a part of DNA sequencing. This project is for reading DNA sequences directly from DNA sequencing gel autoradiographs within a general purpose image processing system.

The DNA sequence reading software is developed based on the waterfall software development approach combined with exploratory programming. Requirement analysis, software design, detailed design, implementation, system testing and maintenance are the basic development stages. The feedback from implementation and system testing to detailed design is much stronger in image processing than a lot of other software development.

After an image is captured from a gel autoradiograph, the background of the image is normalised and the contrast is enhanced. The captured image consists several lane sets of

bands. Each of the lane set represents one part of a DNA sequence. The lane sets are separated automatically into subimages to be read individually. The gap lines between the lane sets are detected for separation. The geometric distortions are corrected by finding the boundaries of the lane set in the subimage. The left boundary of the lane set is used to straighten lane set and the right boundary is used to warp the lane set into a standard width. If separation of the lane sets or geometry correction is unsuccessful by automatical processing, manual selection is used. After the band features are enhanced, the individual bands are extracted and the positions of the bands are determined. The band positions are then converted into the order of the DNA sequence. Different part of a sequence from subsequences are merged into a longer sequence.

In most of the cases, the individual lane sets in a captured image are able to be separated automatically. Manual processing is necessary to handle the cases where the lane sets are too close.

The system may reach an accuracy of 98% if the bands are clear. Manual checking and correcting the detected bands helps to obtain a reliable sequence. If a lane set on the autoradiograph is indistinct or bands are too close it may reduce the accuracy, in extreme cases to the point where it is unreadable. For a 512×512 image captured from a gel autoradiograph, preprocessing takes 90 seconds, processing each subimage takes 40 seconds on a 33Hz 486 PC. If processing a 430×350 mm autoradiograph with 16 lane sets, assuming 6 images are required, it takes about 40 minutes.

Preface

As I complete my study for my master's degree, I want to call to mind my mother and father and sisters, and my whole family who have encouraged and supported me so much over the past years. I am very grateful to my family, who have helped me to persist in my studies when I often felt too old to continue. I feel very relieved and thankful that after so many years of what seems like wasted time and many struggles I have now completed my work.

When I was a small girl I loved studying very much, and was inspired by my parents, and by my sisters. However, circumstances were very difficult from my youth onwards, as a result of the environment at that time.

My parents had no opportunity to get an advanced education when they grew up in China. So they placed their hope on their children. Indeed their greatest desire was that their children would get a good education. As their daughters, we sought not to disappoint their expectation. At school we all got high grades in our work. My eldest sister, Guizhen, and my second sister, Guimei, passed the entrance examinations and were both admitted to university. My third sister, Guie, delayed going to middle school because she had been competing in wushu, a sport in which she became outstandingly

successful and the champion for the whole of China. After middle school, she too went to university. I am the fourth in the family, and I was admitted to one of the best middle schools in Qingdao city after taking the entrance examinations.

My sisters set a fine example for me, inspiring me subconsciously to go on from school to university, and then after my undergraduate study to undertake postgraduate research, and then do research or teach in a university. It all seemed so straight-forward. But the Cultural Revolution in China commenced just as I was completing my three years of study at middle school. So overnight the dream of studying in university evaporated. High school study is for a period of three years between middle school and tertiary study in China, and I felt sure I could go there. However, because my father had once been a capitalist I was not permitted to attend, even though my school marks were always excellent. When I discovered that I would not be permitted to study in high school I was broken-hearted. In my dreams I would imagine that I was at school, and when I woke up, I would cry because I would never be allowed to take a seat in the lecture theatre.

Later, I went to the countryside and worked in a factory. All the "golden time" of my youth had come to nothing. It was ten years later when once again people were admitted to university by sitting the entrance examinations. However, I had attended no classes at high school level, while others had attended for three years. My sisters encouraged me to take the entrance exams anyway, and with their help and my own study I got very good marks much to my surprise. Even then I encountered prejudices against admitting me to university, and I almost missed out on gaining entrance. But I was stubborn, and I clung onto the gleam of hope and refused to succumb to the opposition. So finally I entered my university career, ten years late. My university study in China was the essential preparation to overseas study.

New Zealand is the first foreign country I have ever visited. As a mature student, I found study overseas much more difficult than for others in the same course. Language is one problem, but I have also experienced a serious ailment in my back, and although I do not look ill, I have constantly had to struggle with it. Nobody else has any idea just how acute this little problem has been. I spent a great deal of time trying to control the pain, so that I could concentrate better on my work. The suffering reduced my ability to work and to live, and this made me lose confidence. The worry and distress made my condition worse, and several times I almost abandoned my study. Yet my family's desire that I succeed, and my own long-cherished desire from when I was a child, kept me going, preventing me from giving up. It may have been very hard, but I am determined to go forward and upward.

God has taken care of me. Just before completing my MSc I again almost lost confidence about my future direction. Just at this point the pain in my back began to reduce. So the hope of what lies in the future has returned to me. The desire to climb forward and upward has returned to my heart. I am still my stubborn self, and I will not give up on my aspirations.

During the difficult experiences of my life, I have always had the complete support of my whole family to get me over dangers and difficulties. My mother has always been my guide through life. Whenever difficulty or danger has come near, I have always remembered when I was a child lying in my cradle. She would pick me up and place me securely in her arms. In a way this is what she still does when I am in difficulties. She has always protected and helped her children and encouraged them to go forward and upward. I wish my mother good health, much happiness and long life.

During the two years when I was working on my masters in New Zealand, two members of my family have died, my honoured father and my third sister Guie. They were so far away across the oceans that they could not farewell me before they went to heaven. The pictures of two roses in the Introduction of my thesis are a special tribute to my father and my sister Guie.

When I left home to come to New Zealand, my father wept. He had a premonition that he would never see his youngest daughter again. But I do not believe that my father and sister Guie have left the world. I dreamed about my father last night. I know from this that my father is still alive. So I have a good reason to do what my family want and what they hope for me. I can feel that God is with me, and peace and happiness are present with me, for that is what my sister Guie prayed for and wished as blessing for me in her will. So I want my father, my sister Guie and all my family to smile upon me as I live here.

This thesis is therefore a gift to my father and mother as a thanksgiving for the way they brought me up, and also to my sisters in thanksgiving for their care and love. It is also a memorial to my father and my sister Guie.

Acknowledgments

My all family, my parents, my sisters, my brother-inlaws, my nieces and my nephews, have encouraged and supported me over the past years when I study in NZ. Many thanks to everyone in my family.

Many thanks to my supervisor Dr Donald Bailey, for his valuable guidance throughout this project, for his helping me to learn image analysis techniques, for his patience in trying to understand my Chinese way of thinking and talking, especially when checking my thesis and conference papers, for his encouragement and support when I felt like giving up since my father and sister died.

I would like to thank Dr Nick Ellison for providing invaluable technical information on DNA sequencing and for the provision of the autoradiographs used to develop the software.

I would also like to thank Professor Mark Apperley and Dr John Hudson for their encouragement and support and to thank Terry Cunniffe for being a nice neighbour and helping set up memorial for my father with David's help.

CONTENTS

Chapter 1

Introduction

Digital image processing has been a rapidly evolving field during the last 30 years with a growing range of applications in a broad spectrum of science and engineering disciplines [Jain, 1989]. This growth is coupled with improvements in the processing speed, image display and storage capabilities of computers and cost effectiveness of the related signal processing devices and computers [Pratt, 1978]. Image processing is a broad subject of interdisciplinary study and research in such diverse fields as computer and information science, statistics, physics, astronomy, chemistry, biology, psychology, medicine, geology, engineering and so on [Bailey, 1985].

DNA sequence reading is one application of image processing in fundamental genetic and cellular analysis. Genetic and cellular analysis is an important part of biological, agricultural and medical research [Bodmer, 1987]. The goal of this project is to incorporate automatic DNA sequence reading capability from a gel autoradiograph within a general purpose image processing system.

The genetic information of a living organism is encoded by the DNA contained within every living cell of that organism. DNA sequences are a representation of the genetic structure of DNA molecules. After the DNA reactions are run on an electrophoresis gel

the DNA sequence can be read from the gel autoradiograph. A DNA sequence reading system is developed for reading sequences directly from DNA sequencing gel autoradiographs by image processing techniques. A number of subsequences in an image are captured from a gel autoradiograph. Each subsequence must be read separately. Most of the subsequences may be separated automatically. If necessary, manual processing may be used to separate the subsequences. The subsequence boundaries are extracted and are used to correct for geometric distortions. Individual bands are extracted and the band positions are detected. The order of the DNA sequence is then determined from the sequence of band positions. Different parts of a sequence from different subimages are merged into a longer sequence. The algorithms of the DNA sequence reading system are developed and implemented using VIPS (Vision Image Processing System), version 4.1, which currently runs on an PC under Windows, an Apple Macintosh and a DEC Micro VAX.

Chapter 2 includes a brief survey of image processing and introduces the Vision Image Processing System (VIPS). Some of the image processing concepts are explained which are used in the following chapters on DNA sequence reading software development. These are image acquisition, feature enhancement, linear convolutional filters, intensity histograms, image segmentation, thresholding and line profiles. VIPS is used as the software development and implementation environment for this project. The hardware components required by VIPS and software features of VIPS are described.

In chapter 3, the DNA sequencing application is described more fully and an image processing software development model, based on the waterfall software development approach combined with exploratory programming, is presented. Each stage of the model (requirement analysis, software design, detailed design and implementation, system testing and maintenance) is described as it relates to DNA sequence reading system development. A general model for image processing and the final system function module structure are given in section 3.4. The control module algorithms of DNA sequence reading system are described in section 3.5.

Chapter 4 describes the algorithms for the image processing module of the DNA sequence reading system in detail. An image must be acquired from a DNA sequence gel autoradiograph. Contrast enhancement increases the faint bands in the DNA sequence images. Several subsequences may be in the same captured image and each subsequence must be read separately. The gap lines between subsequences are detected, which are used to obtain separate subimages for each subsequence. Geometric distortions often occur on gel autoradiographs. In order to successfully process most of

the gel autoradiographs, geometry correction is necessary. The subsequence boundaries are extracted and are used to correct for geometric distortions. The left boundary is used to shear the image to straighten left side and the right boundary is used to trapezoid warp the right side. The individual bands are extracted and the positions of the bands are then detected. The band positions are sorted into a list and then converted into a DNA sequence. Different parts of a sequence are merged into a longer sequence.

Chapter 5 discusses the results of the DNA sequence reading system. Automatic separation of lane sets and geometry warping are successful on most of the images captured. If the boundaries of a band lane set are not clear, manual selection of the lane set and manual geometry correction are used. The accuracy of the system depends on the clarity of the bands on the gel autoradiograph. If the bands are clear, the accuracy of automatic processing may reach 98%. The output is checked and any errors may be corrected manually to obtain an acceptable accuracy. Timing is carried out on a 33MHz 486 PC. Automatically processing a 430 × 350 mm autoradiograph with 16 lane sets (see Figure 3.1-1) takes about 40 minutes.

Chapter 6 summarises the system and the thesis and gives suggestions for future work.

The mathematical expression of the VIPS operations used in this thesis are summarised in Appendix I. Appendix II gives VIPS programs for the DNA sequence reading system. Several new VIPS commands were developed for DNA sequence reading: **STRAIGHTEN**, **PARALL**, **SEQUENCE**, **POINTS**, **SORT** and **JOIN**. The C programs for these commands are given in Appendix III.

During the research for the project the following papers were published:

Fan B., Bailey D.G. and Hudson J., Image processing in DNA Sequence Reading, 7th NZ Image Processing Workshop, pp 117-122, Christchuch NZ, (8, 1992).

Fan B. and Bailey D.G.,Algorithms for DNA Sequence Reading by Image Processing, NZ Computer Science Research Students' Conference, pp 109-116, Hamilton NZ, (10, 1992).

Fan B. and Bailey D.G., Algorithms for DNA Sequence Reading by Image Processing, NZ Journal of Computing, pp 47-56, Palmerston North NZ, (5, 1993).

Fan B. and Bailey D.G., Preprocessing Algorithms for Automatic DNA Sequence Reading, TENCON' 93 Computing, pp 998-1001, Beijing China, (10, 1993).