Title
DNA sequencing by denaturation

Author
Chen, Ying-Ja

Publication Date
2009

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

DNA Sequencing By Denaturation

A Dissertation submitted in partial satisfaction of the requirements for the degree

Doctor of Philosophy

in

Bioengineering

by

Ying-Ja Chen

Committee in charge:

Professor Xiaohua Huang, Chair
Professor Michael J. Heller
Professor Sungho Jin
Professor Bing Ren
Professor Shankar Subramaniam

2008

The Dissertation of Ying-Ja Chen is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2008

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to first thank my advisor, Dr. Xiaohua Huang, for providing a stress-free environment for doing research allowing creativity and innovation to flow out of free-thinking. His guidance and training has been extremely valuable for my growth of becoming an independent researcher. I am grateful that he has been fully supportive even when mistakes were made and productivity was low, without which I would not have been able to learn to appreciate science and experimentation with a deep understanding and complete the work presented in this dissertation.

I would like to thank my committee members, Dr. Shankar Subramaniam, Dr. Michael Heller, Dr. Bing Ren, and Dr. Sungho Jin, who were never hesitant in providing valuable discussions not only on my project but also suggestions for the graduate career in general.

Ph.D. is a long and lonely road. I am lucky to have incredibly friendly and intelligent labmates to share crazy or ordinary ideas, useful or trivial knowledge, novel or well-established laboratory techniques, as well as all the up and downs encountered as a graduate student. Vivi Talstad taught me several techniques in the lab. Kristopher Barbee was the one that brought in a social atmosphere to the lab in the early days along with our undergraduate students, Quang Pham, Simpledeep Banipal, Victoria Win, Julianna Yeung, Ronnie Chen, and Jessie Wong. Several experimental breakthroughs were fostered from the ideas he contributed. Eric Roller has assisted in

the construction of our imaging system and took care of most of the programming of many tedious parts and devices. Without his hard work, I would not have been able to focus on conducting the necessary experiments for the SBD project. Aric Joneja was always there to share my excitements and frustrations and remind me that I have made progress. The four of us were active in many aspects although our journal club has only lasted a month. Erin McElfish and Nora Theilacker, who joined the lab later, brought in more diversity while keeping the vibrant and harmonious spirit in the lab.

Besides my labmates, there were many friends who were in the same journey to share the graduate school experience with me. Zac Hsieh, Iona Chen, KJ Yang, Michelle Lee, and Chi-Hui Cheng, who listened to my emotional discharges weekly; Arthur Hsu, Jessica Chang, Sky Yu, Yi-Hsien Su, Chia-ho Tsai, Hsiu-Chin Lin, Sam Chen, Tingfan Wu, and other fellow iHutters who were willing to speak and listen about science at least once a month at our seminar club, iHut; Susan Su, who convinced me by action that graduation is not a dream, along with other bioengineers, Ramses Agustin, Ian Lian, Angela Young, Linda Chang, Albert Hsiao, and Eran Rosine; Ivy Tseng, Hou-Shin Chen, Janet Chen, and Po-Lin Lai, who although were at other schools shared each stage of a graduate student life with me; and all the Taiwanese graduate students at UCSD, you have made my life in San Diego colorful and exciting with activities including surfing/boogie boarding, softball, and volleyball, which greatly compensated for the lonely research life.

Lastly, I would like to express the deepest appreciation to my family for their constant support. My mom, Bon-chu Chung, has acted as my second advisor holding

weekly progress reports sometimes more often than my real advisor even though she was 16 hours of time-zones away. My dad, my grandparents, and my aunts and uncles have also been great support allowing me to be free of worries about life and provided occasional advice for my graduate studies. My family was my inspiration and motivation for getting this degree. Their constant encouragement has helped me build up my confidence and the resilience to complete the work no matter how many frustrations I had to face. Finally, I would like to thank my husband, Sheng-hong Chen. He was the one who was always by my side without asking for a reason. Almost like my third advisor, he was always curious about my research and wanted to discuss in more detail. Occasionally, he even does experiments for me to help figure out the problems. I cannot be more thankful to receive so much support and affection from him and my family.

Chapters 2, 4 and 5, in part, are rearrangements of the material as it appears in "DNA sequencing by denaturation: Principle and thermodynamic simulations", Chen, Ying-Ja and Huang, Xiaohua, *Analytical Biochemistry*, 384(1): 170-179 (2009). The dissertation author was the primary investigator and author of this paper.

VITA

2002 Bachelor of Science, Electrical Engineering, National Taiwan University

2005 Master of Science, Bioengineering, University of California, San Diego

2008 Doctor of Philosophy, Bioengineering, University of California, San Diego

PUBLICATIONS

Li L-A, Chiang F-L, Chen J-C, Hsu N-C, Chen Y-J, Chung B-c, "Function of steroidogenic factor 1 domains in nuclear localization, transactivation, and interaction with transcription factor TFIIB and c-Jun." *Molecular Endocrinololgy* 13: 1588-1598 (1999).

Chen Y-J, Chen Y-C, Lee C-H, Wang J, "High-aspect-ratio sub-diffraction-limit objects fabricated with two-photon-absorption photopolymerization." *Proceedings of the Conference on Lasers and Electro-Optics*, 1: 252-253, Long Beach, CA (2002).

Chen, Y-J and Huang, X, "DNA sequencing by denaturation: Principle and thermodynamic simulations", *Analytical Biochemistry*, 384(1): 170-179, (2009).

Chen, Y-J, and Huang, X, "DNA sequencing by denaturation: Experimental proof of principle", *in preparation*.

ABSTRACT OF THE DISSERTATION

DNA Sequencing By Denaturation

by

Ying-Ja Chen

Doctor of Philosophy

University of California, San Diego, 2008

Professor Xiaohua Huang, Chair

Genome sequencing technologies are in high demand for applications such as gene expressions, the studies of complex diseases, and personalized medicine. In this thesis, I present my work on a new DNA sequencing method called sequencing by denaturation (SBD). A Sanger sequencing reaction is performed on the templates on a surface to generate a ladder of DNA fragments randomly terminated by fluorescently-labeled dideoxyribonucleotides. The labeled DNA fragments are sequentially denatured and the process is monitored by measuring the change in fluorescence

intensities from the surface. By analyzing the denaturation profiles, the base sequences of the templates can be determined in a massive parallel manner.

Using thermodynamic principles, we simulated the denaturation profiles of a series of oligonucleotides ranging from 12 to 32 bases and developed a base-calling algorithm to decode the sequences. These simulations demonstrate that up to 20 bases from a DNA molecule can be sequenced by SBD.

The instrumentation for performing SBD has been constructed by integrating fluorescence imaging, temperature control, and fluidics onto a single device through a custom-made biochemical reaction chamber. This system is fully automated and its performance has been characterized. It can be useful for many applications utilizing high-throughput fluorescence imaging.

Experimental proof of concept for SBD was established by measuring denaturation curves of 6 fluorescently-labeled oligonucleotides hybridized to a common template on the surface. The melting temperature of each oligonucleotide was distinguished correctly. These results demonstrate that experimental measurements of denaturation profiles can be performed on a surface with single-base resolution, which proves the feasibility of SBD.

The throughput of the system was calculated. It can potentially allow up to 200 million DNA templates to be sequenced within 7~20 hours producing 4.2 billion base sequences in a single run. The cost to sequence a mammalian genome is estimated to be about 1000 US dollars. The potential limitations and methods for further

improvement of SBD are discussed. With its high throughput and simplicity, SBD could potentially result in a significant increase in speed and reduction in cost in large-scale genome re-sequencing.

Chapter 1 Introduction

1.1 Genome sequencing: significance and applications

1.1.1 Applications in biomedical research

In 2001, the consensus sequence of the human genome was decoded and this ushered in the post-genomics era [1, 2], when many fields such as functional genomics, comparative genomics, proteomics and systems biology emerged. However, the demand for genome sequences from many individuals and populations has never been greater [1, 2]. In evolutionary comparative genomics, genomes from many different species need to be sequenced in order to be aligned and compared. For the diagnostics of complex diseases or traits, haplotypes need to be identified. This requires the genome sequences from many individuals with and without the disease. In cancer research, the genomes of cancer cells versus normal cells will enable the identification of genetic etiologies of cancer [3]. Furthermore, genome sequencing technology has a variety of applications in many forms of diagnostics.

1.1.2 Traditional sequencing technology and its limitations

Traditional sequencing technology implements the Sanger dideoxy sequencing reaction [4], where dideoxynucleotides as well as deoxynucleotides are added to a polymerase reaction. As the dideoxynucleotides are incorporated randomly, each of its products terminates at specific bases corresponding to the

dideoxynucleotide species. These DNA strands can be separated by gel electrophoresis, and the sequence could then be read out.

Traditional Sanger sequencing method was improved by incorporating fluorescent detection [5] and miniaturizing the gel electrophoresis into a capillary device [6]. Four different fluorescent dyes could be incorporated to the sequencing primer corresponding to each of the four dideoxy reactions. These products could all be run in a single lane of a gel, and a fluorescence detector could be used to read out the sequence based on the corresponding color of fluorescence. For miniaturization, the separation matrix equivalent to the gel is loaded into a capillary so that smaller volume, higher voltage and faster separation speed could be achieved.

Despite the many advances that allowed the sequencing of the human genome, current technology is still expensive and slow for the sequencing of many human individuals or populations. This is mainly due to two factors. First, the sample preparation pipeline for sequencing is labor-intensive and requires the individual manipulation of every clone. Second, capillary gel electrophoresis has a limited throughput. For these reasons, new technologies must be developed to overcome the technical barrier.

1.2 New genome sequencing technologies

1.2.1 Genome sequencing pipeline

The sequencing of a genome entails several steps summarized in Figure 1. First, genomic DNA is isolated from cells and sheared into millions of fragments. A genomic DNA library is constructed by cloning these DNA fragments into vectors. Next, the library of clones are separated and amplified by polymerase chain reaction (PCR). Then, these amplified clones are sequenced one by one in an automated DNA sequencer. After decoding the sequence of each individual fragment, they are assembled into the genome sequence by aligning the overlapping sequences with very sophisticated computation algorithms.

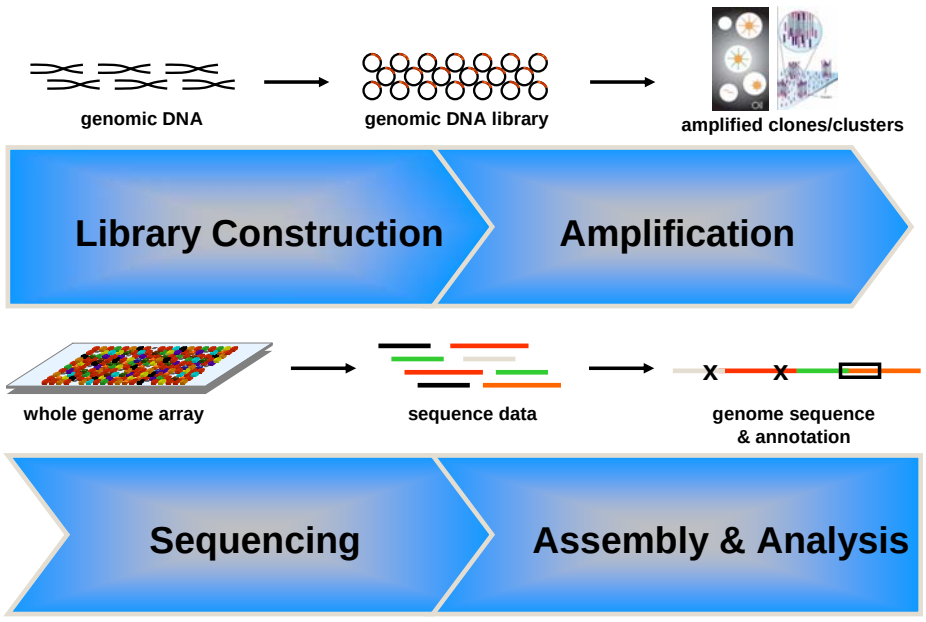


Figure 1. Genome sequencing pipeline. Genomic DNA is first fragmented and constructed into a genomic DNA library. Then the library is amplified either on the surface or later separated onto the surface for sequencing. After sequencing, the sequences are aligned to the reference sequence or assembled.

Several sequencing platforms have been developed for DNA sequencing with a faster speed and lower cost. Some or all of the steps described above are improved in some way in these DNA sequencing platforms. In genomic library construction, cloning into bacteria is eliminated and replaced with the attachment of universal adaptors which will facilitate the downstream amplification and sequencing steps. Genomic DNA amplification is performed in micro-compartments either in water-in-oil emulsions or on spatially separated surfaces so that the amplification of millions of clones can be achieved in one reaction. Similarly, several new sequencing chemistries are being developed so that sequencing can be performed in parallel on a solid surface. Finally, bioinformaticists are developing new DNA assembly algorithms for assembly from sequence reads that are much shorter than the traditional ones. The details of some of the promising methods are explained in the sections below.

1.2.2 Genomic library construction and amplification methods

In order to amplify the whole genomic DNA library in a high-throughput fashion, each individual clone must be separated and amplified from a single molecule. The amplification method must demonstrate the capability for parallel processing. Many methods have been developed to amplify genomic DNA library by PCR.

One of the most common methods used in both the 454 and the Applied Biosystems systems is the emulsion PCR method developed by Vogelstein's

group [7]. In emulsion PCR, primers are attached to the surface of microbeads. PCR reagents, target templates, and the primer-containing beads are mixed in a water-in-oil emulsion. These micro-emulsions become compartments separating each individual clone while serving as micro-reaction chambers for the PCR amplification. The DNA fragments are cloned onto the beads, which then could be spread across a surface for sequencing. While achieving highly efficient amplification in a multiplexed fashion, amplification of different clones can be biased significantly due to the inhomogeneous sizes of emulsions.

Another method developed by the Church group is called polony which stands for PCR colony [8]. Template DNA and PCR reagents are poured onto a glass surface with polyacrylamide matrix. The polyacrylamide film restricts the diffusion of DNA into a localized area so that each individual template could be PCR amplified without any crosstalk with other clones. This allows millions of clones to be amplified on one microscope slide in one reaction. In addition, by using primers with a 5' acrydite attached, the products could be covalently linked to the polyacrylamide matrix and allow for further manipulation such as in situ sequencing [9]. The Solexa/Illumina system employs a similar method called bridge PCR where both primers are immobilized on a solid surface so that the products are restricted to the local area where the template was first hybridized [10]. These methods have the advantage of uniform amplification. However, because the amplification depends on the products folding down to hybridize with the next primer, amplification efficiency is lower than emulsion PCR.

1.2.3 Massively parallel DNA sequencing methods

Many efforts have been devoted to miniaturization and parallel processing for high-throughput DNA sequencing [1, 2, 11]. Miniaturized gel electrophoresis methods are the most straightforward because it is based on the current sequencing principle. Richard Mathies' group has fabricated 384 capillary gels on a wafer [12], where samples are loaded from the outer rim of the wafer, and electrophoresis is run towards the center of the wafer. A rotary confocal fluorescence scanner is used for detection. However, because the required spatial separation limits the number of micro-capillary gels that can be fabricated on a wafer, this method could only improve the sequencing throughput by several folds, which is still too low for sequencing a mammalian genome.

In order to improve the technology by another 100 times to enable routine sequencing of genomes in medical centers, other principles have been developed to enable sequencing on a chip or microfluidic device, which have higher throughput in nature. One popular solution is to use DNA polymerase to incorporate a single base at a time onto the template sequence, decipher that base, and then repeat the cycle until the maximum read length is reached. The most common of these methods, often referred to as sequencing by synthesis (SBS) or sequencing by extension (SBE), uses fluorescently-labeled dideoxynucleotides with a blocking group at the 3'-OH to prevent the next dideoxynucleotide from being incorporated before the current base is decoded [9, 13]. Significant progress has been made in engineering polymerases that can incorporate modified

nucleotides [13] and engineering nucleotides with cleavable reversible terminators [14, 15]. The Illumina/Solexa 1G Analyzer utilizes this technology [10]. Due to the low incorporation efficiency of modified nucleotides by DNA polymerase and the incomplete cleavage of reversible terminator groups resulting in fewer accessible templates and higher background fluorescence, the read length of SBS has been limited to 35 bases. This problem has been partially alleviated by a SBS-Sanger hybrid method which uses a mix of nucleotides with reversible terminators and natural bases with cleavable fluorescently-labeled dideoxynucleotides. The reversible terminators are smaller and easier for the DNA polymerase to incorporate while the cleavable fluorescently-labeled dideoxynucleotides can provide enough signals for detection.

One method that bypasses those challenges in SBS is pyrosequencing, which detects the level of pyrophosphate group released as each natural nucleotide species is incorporated [16, 17]. That is done by coupling two enzymes to the reaction: sulfurylase, which uses pyrophosphate as a substrate to generate ATP, and luciferase, which in the presence of ATP converts luciferin to oxyluciferin resulting in the production of photons. This method has been used by 454 Life Sciences to build their DNA sequencer such as the Roche/454 GS-FLX [16]. A honeycomb shaped fiber optic plate was used for detection of the light emitted by luciferase. Challenges such as sequencing homopolymers have been addressed by controlled nucleotide concentration and quantifying the light emission. This was the first next-generation sequencer to become commercially

available and the only one currently with a read length of greater than 35 bases. However, in order for many templates to be sequenced on one surface, the pyrophosphate must be detected locally and in real-time before they diffuse away and contaminate the detection of neighboring templates. Thus, its throughput is still limited by the detection method and the inability to further scale down the size of the picotiter reactors.

Besides SBS, sequencing by hybridization (SBH) is another method of high-throughput nature based on microarray technologies. The majority of these technologies focus on resequencing rather than *de novo* sequencing. On resequencing microarrays developed by Affymetrix and Perlegen, 25 bp oligonucleotide probes are synthesized on the resequencing microarray [18, 19]. For each position interrogated, four probes are made with the same sequence with the exception of the center one which varies among A, C, G, and T. The sequencing template is labeled and hybridized to these arrays and the sequence is read out by finding the highest relative fluorescence intensity in each feature set. Another method reported recently immobilizes the templates on a solid surface and hybridizes all possible pentamers to probe the templates [20]. Algorithms are used to decode the sequence. SBH methods eliminate complications of enzyme efficiencies but suffer high error rates due to cross-hybridization of template to mismatch probes.

The Applied Biosystems' (recently merged with Invitrogen) SOLiD system is based on hybridization and ligation. In sequencing by ligation (SBL)

[21-23], all possible nonamers with a fluorescent tag encoding for only the center base are incubated with the template to allow the correct probe to hybridize to the template adjacent to the anchor primer. Then the nonamer is ligated to the primer in the presence of DNA ligase. The base in the center is decoded by fluorescence measurement. Now another nonamer could be hybridized and ligated to sequence the base that is 9 bases downstream of the previously sequenced base. After going through the whole sequence, the entire strand is stripped, and a new primer with an offset is attached to sequence another set of bases that are 9 bases apart. After cycling through 9 different primers, the whole sequence is decoded. This method eliminates the need for polymerases that could incorporate non-native nucleotides as in SBS. However, ligase does not discriminate mismatch nucleotides from matches as well as DNA polymerase does, which increases the error rate and decreases the efficiency. SBL accuracy and efficiency has been improved by cleaving away the last few bases per cycle to sequence every 5 bases rather than 9 and by using a two-base color encoding scheme.

These cycling methods have a common limitation on the synchronization of reaction within each clone to be sequenced. If a few strands were not incorporated in the previous cycle, they could still be incorporated in the next cycle and give a false signal. The strands become totally out of synchronization after several cycles and the sequencing limit is reached. However, by monitoring only single molecules, this would not be a problem. Many groups are developing detection methods to use the cycling reaction methods to sequence single

molecules. There are various ways to detect the signal, such as total internal reflection fluorescence (TIRF) microscopy, fluorescence resonance energy transfer (FRET) [24], or the nanofabricated zero-mode waveguide being developed at Pacific Biosciences that only allows fluorescence to be detected when it is at the polymerase site [25-27]. These methods require high sensitivity imaging and the elimination of background fluorescence. In order to prevent the fluorescent molecules from photobleaching before detection, oxygen scavenging, free radical scavenging and triplet quenching components to the sequencing buffer are added into the buffer for imaging [28, 29]. Although cost is reduced by its high throughput and the elimination of clonal amplification, sequencing speed is still limited.

An alternative to fluorescent detection is to use electric signal, such as current detection. In nanopore sequencing, a single DNA strand is pulled through a nanopore whose size only allows a single DNA strand to go through. As the DNA strand goes through and blocks ionic current, there will be a current change that can be measured to determine the base that is crossing the nanopore [30, 31]. Many technical difficulties need to be solved for this method to work, such as methods to determine the identity of the base. Although the concept is promising, DNA sequencing with nanopores has not been demonstrated.

1.3 Principle of DNA sequencing by denaturation

Denaturation is the reverse process of hybridization. A DNA sequencing method based on hybridization called sequencing by hybridization (SBH) was

proposed many years ago as a high throughput sequencing method [32-34]. In SBH, a target sequence is interrogated by the differential hybridization of short perfectly complementary probes [18, 19, 32-36]. SBH has been difficult to implement primarily due to the complexity associated with the SBH process, particularly the cross-hybridizations of probes to incorrect but similar sequences in the context of a complex mixture of probes and target sequences. In contrast to hybridization, denaturation is a simple and relatively slow process which strictly depends on the thermodynamic properties of the DNA molecules [37]. Denaturation does not require the initial collision of single-stranded DNA species and thus is free of the complexity associated with the hybridization kinetics. It is also unaffected by cross-hybridization of mismatched probes to the target DNA. Therefore, we reason that denaturation could be used for DNA sequencing by performing melting curve analysis.

The basic principle of SBD is illustrated in Figure 2. As shown in Figure 2, first, a standard Sanger sequencing reaction is performed using fluorescently-labeled dideoxynucleotides on the templates immobilized on a surface. Instead of being resolved by gel electrophoresis, these randomly terminated DNA fragments are sequentially denatured and washed away by applying a denaturation force such as an increase in temperature or the concentration of a chemical denaturant. As each fluorescently-labeled dideoxy-terminated fragment denatures, the fluorescence in the

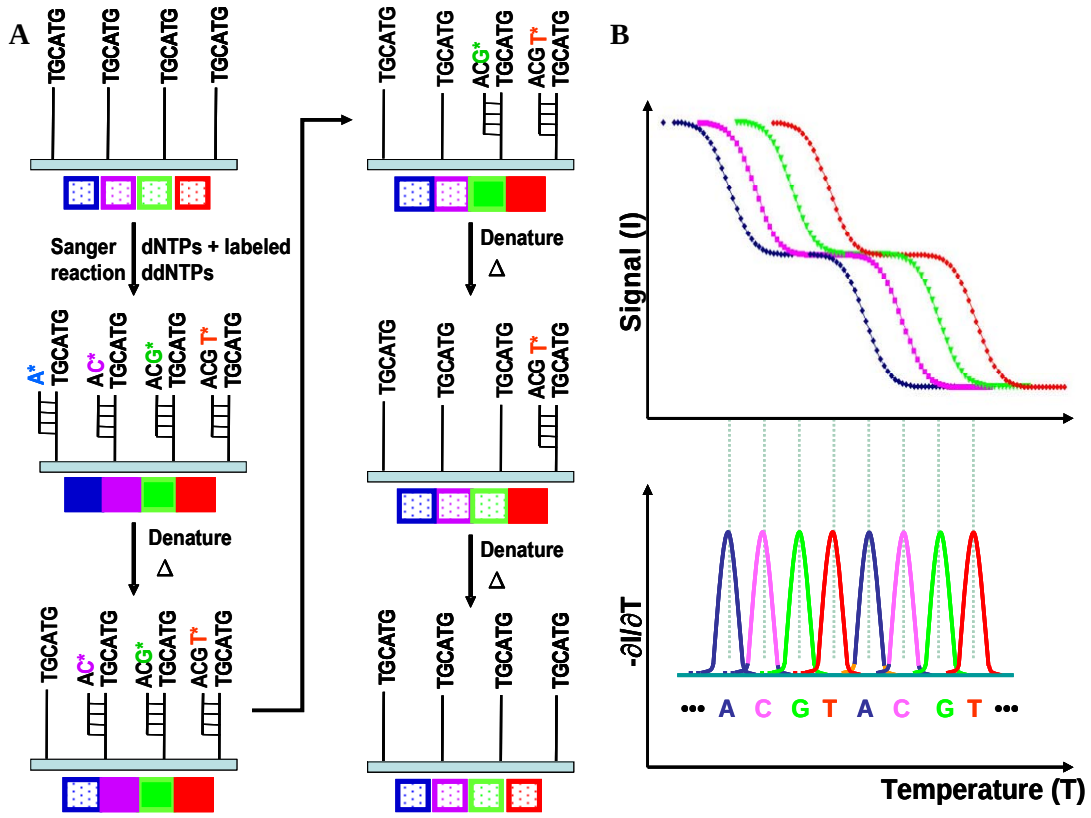


Figure 2. Schematic of SBD. (A) The process of SBD is shown. Amplified DNA templates are immobilized on a surface. Sanger sequencing reaction is performed with deoxyribonucleotides and fluorescently-labeled dideoxyribonucleotides, which generates fluorescently-labeled fragments of various lengths. The square boxes represent the fluorescence observed in each channel. Subsequently, the surface is heated to denature these fragments. As each fragment denatures, the fluorescence in the corresponding channel decreases. In this illustration, first the blue channel representing the base A is denatured, followed by the violet channel representing C, the green channel representing G, and finally the red channel representing T. **(B)** The fluorescent signal of the sequencing of a DNA template with 8 bases is shown. In the top graph, the signal intensity in each channel is shown to decrease as each fragment denatures. The bottom graph shows the negative derivative of the fluorescent signal intensity where the sequence of each peak can be sorted to read out the corresponding bases.

corresponding channel on the surface decreases. The ensemble of these melting curves can be measured by monitoring the fluorescence from the fragments that remain

hybridized to the templates. *A priori*, single-base resolution can be obtained because the melting temperature of a DNA strand of certain number of bases is lower than that with one additional base. By analyzing the fluorescent intensity remaining on the surface, which reflects the denaturation event, the sequence of the template can be determined. The graph of the negative first derivatives of the denaturation curves with respect to temperature looks similar to a conventional electropherogram in Sanger dideoxy sequencing by gel electrophoresis. The sequence can be decoded from the order of the peaks in the graph.

This method eliminates cycling reaction and is the most straightforward among all sequencing methods. It has the advantage of simplicity and is high-throughput in nature. It allows for billions of DNA strands to be sequenced on one chip. Although read length may be limited by the ability to distinguish DNA strands by their denaturation profile, its ultra-high throughput nature could compensate for that.

1.4 Melting curve analysis

Melting curve analysis is the measurement and analysis of the melting of DNA in which double-stranded DNA separates into two single strands. This process occurs under various conditions such as an increase in temperature or chemical denaturant concentrations. The melting or denaturation profile can be monitored by optical techniques such as absorption and fluorescence microscopy. For example, the interactions among stacked bases cause a decrease in UV absorption. Melting of double-stranded DNA at elevated temperature involves breaking the hydrogen bonds

of the base pairs and a decrease of base stacking. This results in an increase in UV absorption – a hyperchromicity, which can be measured by a spectrophotometer [37] .

Melting curve analysis has been used in many applications such as the detection of single nucleotide polymorphisms (SNPs) [38-40]. The melting behavior and melting temperature (T_m) of short oligonucleotides can be predicted quite well by the nearest-neighbor thermodynamic model which assumes that the stability of a DNA duplex depends on the identity and orientation of neighboring base pairs [41-47]. In this study, we have extended this nearest-neighbor model to predict full melting curves of an extensive series of short oligonucleotides at various ionic strengths to provide a theoretical basis for SBD.

1.5 Scope of this dissertation

This dissertation presents the development of a new DNA sequencing method called sequencing by denaturation (SBD), including the theoretical and experimental proof of principle of the method. In Chapter 2, the theoretical basis of SBD is described. Using melting curve analysis and thermodynamic principles, melting curves of short oligonucleotides were simulated. A base-calling algorithm was developed to call the base sequence from the denaturation profiles measured. Upon establishing the proof of concept for SBD, we need an imaging system with the capabilities to perform biochemical reactions in a massively parallel fashion for high-throughput sequencing. In Chapter 3, the construction of a high-speed fluorescent imaging system with integrated temperature control and fluidics is described. This system was constructed

with the primary purpose of performing millions of DNA sequencing reactions in parallel although it may be useful for other applications. Issues regarding SBD using this system were fully characterized. In Chapter 4, the proof of feasibility for SBD was established. Denaturation curves of eight oligonucleotides were first measured in solution to demonstrate that the base-calling algorithm developed earlier could be used to determine the DNA sequence correctly on measured data. Then, the imaging system constructed in Chapter 3 was used to perform experiments necessary to establish the proof of feasibility for SBD. The denaturation profiles of short oligonucleotides were measured on the surface. An image processing protocol was developed to process the acquired images with a computer program. These data were analyzed to demonstrate the feasibility of SBD on our instrument. Finally, the advantages and limitations of SBD are discussed in Chapter 5 and suggestions for future work and improvements are also presented.

Chapter 2 Thermodynamic Principle and Simulations of Sequencing by Denaturation

2.1 Introduction

This section describes the theoretical basis and simulations for SBD. In SBD, Sanger reaction products terminated by fluorescently-labeled dideoxynucleotides are denatured sequentially. These denaturation events are monitored by measuring the fluorescence signal from the molecules that remain hybridized on the surface. Here we use thermodynamic principles to explain why denaturation of Sanger products occurs sequentially and how this phenomenon can be simulated using previously reported thermodynamic data. In this chapter, we simulated the denaturation process of oligonucleotides from 12 to 32 bases in length, the SBD data from those melting curves, and developed an algorithm to analyze such data for base-calling.

2.2 Thermodynamic principle of sequencing by denaturation

In order to simulate the data obtained from SBD, we simulated the individual melting curves and ensemble denaturation profiles of short oligonucleotide series. The denaturation profiles of double-stranded DNA fragments were simulated by calculating the fraction of the DNA that remains hybridized as a function of temperature. The equilibrium of reaction equation for a single denaturation event of a Sanger fragment is the following:

$$D \rightleftharpoons S_1 + S_2$$

$$K_{eq} = \left(\frac{(1-f)^2 C_0}{f} \right)$$

where f is the fraction of DNA that remains hybridized in the double-stranded form, T is temperature, R is the gas constant, and C_0 is the initial concentration of the duplex DNA. This equilibrium constant is related to the Gibbs free energy of the system, which can be written as:

$$\Delta G^0{}_{den} = -RT \ln K_{eq} = -RT \ln \left(\frac{(1-f)^2 C_0}{f} \right) \quad (1)$$

In double-stranded DNA, hydrogen bonds between complementary bases and the pi-stacking of aromatic structure of neighboring bases contributes to a stabilizing energy. Under higher temperatures or in a more stringent environment, these non-covalent bonds break and the more favorable state becomes two complementary single-stranded DNA. For short DNA fragments, one extra base-pair contains 2 or 3 extra hydrogen bonds and additional base-stacking, which are significant forces within the double-stranded DNA structure that keep it in the double-stranded form. Therefore, DNA strands with more base-pairs favor the double-stranded state and have higher free energy for the above equilibrium process. When the temperature is raised, the equation for free energy becomes more negative, which favors the process of breaking double-stranded DNA into single-stranded form. That is the process of "DNA melting" or denaturation. The point where there are equal amounts of double-stranded DNA and single-stranded DNA is the melting temperature (T_m).

The sequence of a short DNA strand contributes greatly to the free energy of its denaturation process. This is because there are 2 hydrogen bonds in A-T base-pairs versus 3 hydrogen bonds in G-C base pairs. Additionally the base-stacking forces depend on the alignment of neighboring base-pairs, which in turn depends on the type of bases present. As a result, the free energy of the denaturation process depends on the base-pairs and their neighboring base-pairs.

Although the free energy and other thermodynamic parameters such as enthalpy and entropy have not been measured for the denaturation process, those parameters for its reverse process, hybridization, have been well studied. Thermodynamic parameters of the hybridization process have been measured and a nearest-neighbor model has been established to predict the free energy and melting temperature of short oligonucleotides. The free energy of denaturation is related to hybridization by the relationship below.

$$\Delta G^{0}{}_{den} = -\Delta G^{0}{}_{hyb} \tag{2}$$

Using this relationship and the nearest-neighbor model established for hybridization, we can predict the melting temperature under the denaturation process of short DNA strands according to their base sequence.

In the nearest-neighbor model, the free energy of a DNA fragment is determined by adding the contributions of each of its nearest-neighbor base-pairs, which are negative parameters. Thus, the free energy of the denaturation of a longer strand is a larger positive number than a shorter strand with the same sequence. As a

result, the shorter strand would have a lower melting temperature than the longer strand with the same sequence.

To calculate the denaturation profile which can be represented by the fraction hybridized *(f)* as a function of temperature *(T)*, we use the nearest-neighbor model [41-47] and the thermodynamic parameters of the standard state enthalpy (ΔH^0) and entropy (ΔS^0) for the nearest-neighbor pairs in the DNA sequence from SantaLucia and colleagues [44]. The reported thermodynamic data of enthalpy (ΔH^0) and entropy (ΔS^0) for the nearest-neighbor pairs are the changes of these thermodynamic values in the binding (hybridization) process at standard state at 37 ºC. Since denaturation is the reverse process of hybridization, the Gibbs free energy change used in our calculation is the negative of the ΔG^0_{hyb} calculated from the reported data. Therefore the ΔG^0_{den} in Equation 1 is related to the ΔH^0 and ΔS^0 reported in the literature by the following:

$$\Delta G^0_{den} = -(\Delta H^0 - T\Delta S^0) \tag{3}$$

By combining this with Equation 1, we have:

$$RT \ln\left(\frac{(1-f)^2 C_0}{f}\right) = \Delta H^0 - T\Delta S^0 \tag{4}$$

By solving Equation 4, f expressed as a function of temperature *(T)* is given by:

$$f(T) = 1 + \frac{1}{2C_0 \exp\left(-\frac{\Delta H^0 - T\Delta S^0}{RT}\right)} \pm \sqrt{\frac{4C_0 \exp\left(-\frac{\Delta H^0 - T\Delta S^0}{RT}\right) + 1}{4{C_0}^2 \exp\left(-\frac{2(\Delta H^0 - T\Delta S^0)}{RT}\right)}} \tag{5}$$

For the solution with the "+" sign, the fraction *f* is greater than one. Since a fraction should not be greater than unity, out of the two possible solutions, the one with "-" sign is reasonable and was chosen as our solution. This is the relationship of fluorescent signal *f* versus temperature *T* determined to plot the melting curve.

Salt concentration has a significant effect on the denaturation process and must be taken into account. Monovalent cations such as sodium ions bind to both single- and double-stranded DNA causing conformational changes that affects the denaturation of DNA. The effect of salt concentration is accounted as an ensemble described by the denaturation reaction below.

$$D \rightleftharpoons S_1 + S_2 + v\,Na^+$$

where *D* represents the double-stranded DNA bound with monovalent counter cations (in this case Na^+), S_1 and S_2 represent the two single-stranded forms, and *v* represents the effective number of the sodium ions released during the denaturation process. The equilibrium constant then becomes:

$$K_{eq} = \frac{[S_1][S_2][Na^+]^v}{[D]} = \frac{(1-f)^2 C_0 [Na^+]^v}{f}$$

f expressed as a function of temperature *(T)* is then given by:

$$f(T) = 1 + \frac{1}{2C_0[Na^+]^v \exp\left(-\frac{\Delta H^0 - T\Delta S^0}{RT}\right)} - \sqrt{\frac{4C_0[Na^+]^v \exp\left(-\frac{\Delta H^0 - T\Delta S^0}{RT}\right)+1}{4C_0{}^2[Na^+]^{2v} \exp\left(-\frac{2(\Delta H^0 - T\Delta S^0)}{RT}\right)}} \qquad (6)$$

v can be approximated by empirical methods described by Owczarzy and colleagues [48]. Since melting temperature (T_m) is defined as the temperature where 50% of the DNA remains hybridized, T_m can be determined from Equation 6 as:

$$T_m = \frac{\Delta H}{-R \ln \frac{2}{C_0} + R v \ln\left[Na^+\right] + \Delta S} \tag{7}$$

In SBD, these melting temperatures are measured as the melting profile of all Sanger fragments through fluorescent imaging. By measuring how much the fluorescence in a channel has decreased, we can determine the denaturation event of one of the Sanger fragments. Because the denaturation events follow a distribution, the entire melting profile is measured and analyzed to determine which fragment has denatured. Finally, the base sequence is read out by determining the order of these denaturation events in different channels.

2.3 Simulations of melting curves

Using the equations developed above, we first simulated individual melting curves of oligonucleotides that are 12 to 32 bases in length. The simulations were conducted with MATLAB. For each oligonucleotide, the ΔH^0 and ΔS^0 were calculated by the summation of all the nearest-neighbor pairs and the correction terms in the DNA sequence using the data reported by SantaLucia and colleagues [44]. From Equation 6, the fraction of DNA remaining in the double-stranded form f was simulated as a vector with each element corresponding to a temperature point in the temperature vector T, which ranges from 0 to 100ºC. The initial concentration of the

oligonucleotides C_0 used was 1 μM. We evaluated various salt concentrations ranging from 10 mM to 1 M and chose 10 mM for the example cases presented herein. The parameter v was approximated by the derivative method described by Owczarzy and colleagues [48]. We performed simulations for 348 oligonucleotides provided in the accuracy benchmark developed by Panjkovich and Melo [49] to compare our predictions of melting temperatures to the previous studies and the experimental values. For each oligonucleotide, the DNA sequence and the salt concentration were varied to obtain the predicted melting temperature. These results were used to confirm that the model provides a similar accuracy as previously reported.

For the 348 sequences provided in the accuracy benchmark, the average error in melting temperature prediction was 3.1ºC, which is similar to the previously reported results. Shown in Figure 3A are the simulated denaturation curves of a series of 20 oligonucleotides each containing a common 12-base primer with the sequence "ATTAAACCTTAA" and additional bases from the first to the 20th bases of the sequence "GTCAGTCAGTCAGTCAGTCA". Figure 3B shows the corresponding negative derivatives of the curves with respect to temperature. It is obvious that a shorter DNA fragment denatures at a lower temperature than one with an additional base pair. The longer the DNA strands are, the sharper the transitions become and the smaller the T_m differences between the neighboring bases. For clarity, the simulation results from a sequence with 4 base types evenly distributed along the sequence are shown. Simulations of the 1000 randomized sequences as described above show that the T_m of the oligonucleotides increases monotonically as additional bases are added

onto the primer sequence. These results demonstrate that in theory single-base resolution could be obtained for oligonucleotides up to 32 bases long.

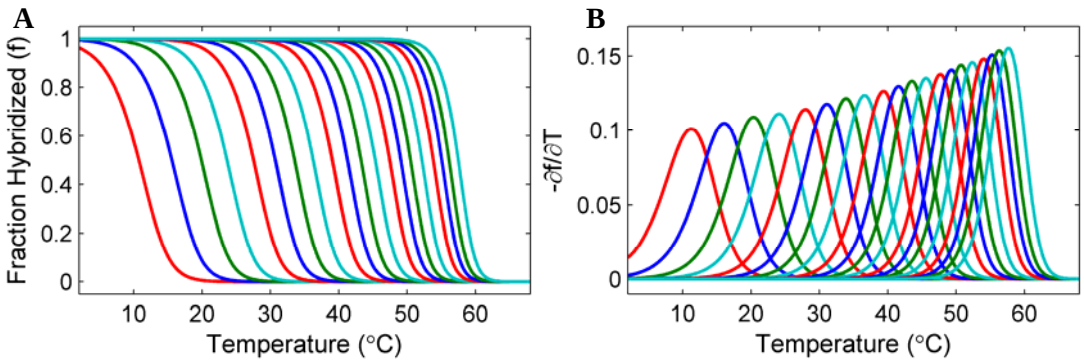


Figure 3. Simulated denaturation curves and their negative first derivatives. (A) Denaturation curves and **(B)** the corresponding negatives of the first derivatives of the curves of a series of 20 oligonucleotides each of which consists of a common 12 base primer with the sequence "ATTAAACCTTAA" and additional bases from the first to the 20th bases of the sequence "GTCAGTCAGTCAGTCAGTCA". The leftmost curve is the simulated profile of the sequence with 13 bases. The rightmost curve is the profile of the full length sequence with 32 bases. A total of 20 curves are shown. As can be seen, the melting temperature increases monotonically as additional bases are added to the sequence.

2.4 Salt effects on melting temperature

As described above, the concentration of monovalent salt influences the denaturation process. The relationship between melting temperature and salt concentration was determined by performing simulations on 1000 random sequences with sodium ion concentrations ranging from 10 mM to 1 M. Each sequence contains a common primer with the sequence ATTAAACCTTAA concatenated with 20 base sequences generated from a random number generator with uniform distribution. For

each of these sequences, the ΔH^0 and ΔS^0 were calculated as each of the 20 bases was added to the primer generating 13- to 32-base long fragments. Then the melting temperature of each fragment was determined using Equation 6. In order to survey the melting temperatures of 13- to 32-base long oligonucleotides, an average over the 1000 randomly generated sequences for each fragment length was calculated. The average melting temperature for the DNA fragments was plotted versus salt concentration.

Figure 4 shows the effect of salt concentration on the melting temperature. The average melting temperatures of oligonucleotides with different lengths are plotted versus salt concentration. Each line represents the average of 1000 oligonucleotides with a common 12-base primer of sequence "ATTAAACCTTAA" and 1 to 20 additional bases. As shown, the melting curves have similar profiles but are shifted towards lower temperatures at lower salt concentrations in a non-linear relationship. This plot provides a comprehensive chart for determining the optimal salt concentration to use for experimental measurements of denaturation profiles in SBD. By using an optimal salt gradient, the observation window can be widened.

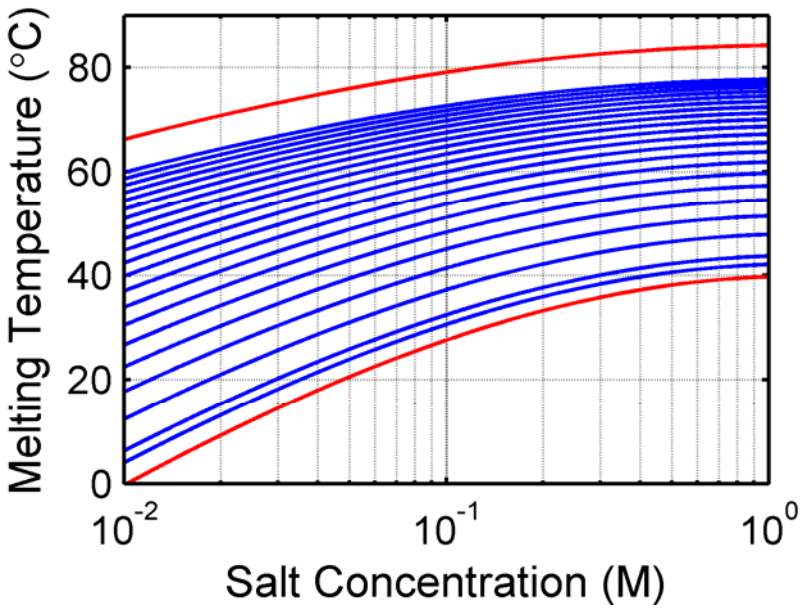


Figure 4. Salt concentration effect on SBD. The average melting temperatures of 1000 oligonucleotides with a common primer sequence of ATTAAACCTTAA and 1 to 20 more additional random bases are plotted versus sodium concentration. Note that the salt concentration is plotted on a logarithmic scale. The bottom red line represents the average melting temperature minus two times standard deviation for the sequences with one base added to the primer. The top red line represents the average melting temperature plus two times standard deviation for the sequences with 20 bases added to the primer. In this figure, the melting temperatures of DNA become lower as the salt concentration is decreased. This plot is useful in determining the optimal salt concentration window for SBD measurements.

2.5 Simulations of sequencing by denaturation

In SBD, a DNA molecule is sequenced by measuring and analyzing the melting curves of the fluorescently-labeled DNA fragments generated by a Sanger dideoxy termination sequencing reaction. The measured fluorescence signal from each color/channel is the sum of the signals from all denaturation curves with sequences ending in the corresponding base type (A, C, G or T). We simulated the fluorescence

intensity profiles accordingly. For a given DNA template, the denaturation curve for each sequence of all of the oligonucleotides, which consists of a common primer and the additional bases along the template, was simulated separately using the methods described in the previous sections. The curves from all the sequences ending at a particular base type were summed to give the overall fluorescence intensity profile for the channel corresponding to that base type. In order to account for noise and variations on sequencing data obtained from real experiments, a Gaussian noise was added to the simulated fluorescence intensity. The noise level was varied from 1% to 10% of the fluorescence signal.

Simulations were performed on one thousand 32-base long oligonucleotide sequences. All the sequences share a 12-base common primer with the sequence ATTAGACCTACG. The other 20 bases in each of the sequences were generated using a random number generation function with a uniform distribution in MATLAB. The salt concentration was fixed at 10 mM. The SBD signal was simulated by the summation of all melting curves ending at the base type corresponding to the fluorescence channel. By analyzing this signal with the base-calling algorithm described in the next section, the DNA sequence is determined. The base-calling accuracy was evaluated by aligning the called sequences to the original sequences. The error rate was defined as the total number of substitutions, insertions, and deletions divided by the number of bases called. The cumulative average error rate was calculated as the percentage of error for a given read length. The base-calling accuracy was evaluated for simulations with 0.1ºC, 0.3ºC, and 0.5ºC sampling

frequencies, and with simulated Gaussian noise values of 0%, 1%, and 5% of the total intensity.

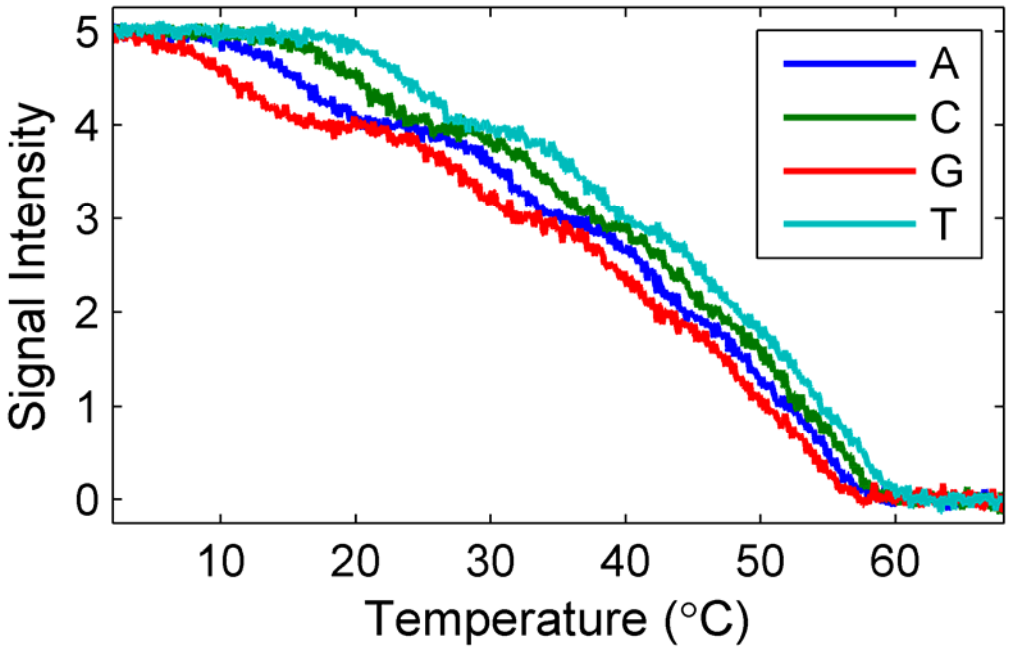


Figure 5. Fluorescent signal of SBD data. Here we show the simulated fluorescent signal for the DNA fragment with sequence "GTCAGTCAGTCAGTCAGTCA". It is the sum of all the contributing melting curves in the corresponding channel. A 5% random Gaussian noise was added.

Figure 5 shows the simulated SBD signal for the same sequence as the one shown in Figure 3. The signal in each channel is the sum of all melting curves that terminate at the corresponding base. The noise level added to simulate measurement error is 5% in this case. These curves show the expected signal of SBD with the appropriate transitions discernible by eye. In the next section, we will show how this data is analyzed to decode the base sequence.

2.6 Base-calling algorithm

The SBD simulated data were smoothed before and after taking the negative derivative curves to generate peaks that mimic the electropherograms generated by traditional Sanger sequencers. Because some peaks overlap and have different width properties from those of traditional electropherograms, an algorithm was developed to find the components of each peak for base-calling.

First, the peaks in each negative derivative curve were found if they were within a defined width and height. Here we assume each Sanger fragment to be equally populated. Because neighboring peaks may overlap, each peak was fit to a sum of Gaussian curves to determine its components. The peak positions determined from the fit correspond to the melting temperature of the component Sanger DNA fragments. In some cases, two adjacent peaks, each containing two or more component peaks, may overlap at the ends of the peaks. A second fit was performed to correct for these cases. This was performed after subtracting the other components based on the first fit so that the fit can be improved. Finally, the peak positions were sorted to decode the DNA sequence.

The algorithm involves five steps which are described in detail as follows.

(1) Take negative derivative from the smoothed SBD signal. The fluorescent intensity signal was smoothed using a moving average filter with window size spanning 3ºC. After the derivative was taken, the signal was smoothed again with the same parameters.

(2) Find all of the peaks. In the negative derivative curves, each peak was identified if its width and height were within reasonable boundaries to capture all of the legitimate peaks. In this step, each peak was characterized by the position, height, start and end of the peak. The peak position was determined by the local maximum in the second derivative of the negative derivative curve. The peak height was the value at the peak position. The start and the end of the peak were defined as the positions where the negative derivative reaches a threshold before and after the peak position, respectively. If two peaks overlap partially, the start or end between those peaks would be the local minimum. These parameters determined the range where each peak was located for fitting in the next step.

(3) Fit each peak to a sum of Gaussians. The melting curves of some Sanger fragments overlap extensively forming a large combined peak. A fit to the sum of Gaussian curves was performed to deconvolve the individual components. The number of Gaussian components in each peak was determined by the area underneath the curve within this peak region. The initial coefficients and lower and upper bounds were chosen in order to achieve adequate Gaussian fits. The initial coefficients of the mean value for the Gaussian fits were equally-spaced values around the peak position within the region. The lower and upper bounds were determined so that the component melting temperatures do not overlap. In order to determine the bounds for the height and the standard deviation of the Gaussian curves, a set of statistical parameters was obtained by performing simulations on 20,000 negative derivatives of the melting profiles of random single DNA fragments fit to

Gaussian curves. The height (A) was determined to be related to the mean position (μ) or melting temperature by the following quadratic equation:

$$A = 1.87 \times 10^{-5}\, \mu^2 + 2.41 \times 10^{-4}\, \mu + 0.1017.$$

The standard deviation (σ) is linearly related to the mean or melting temperature:

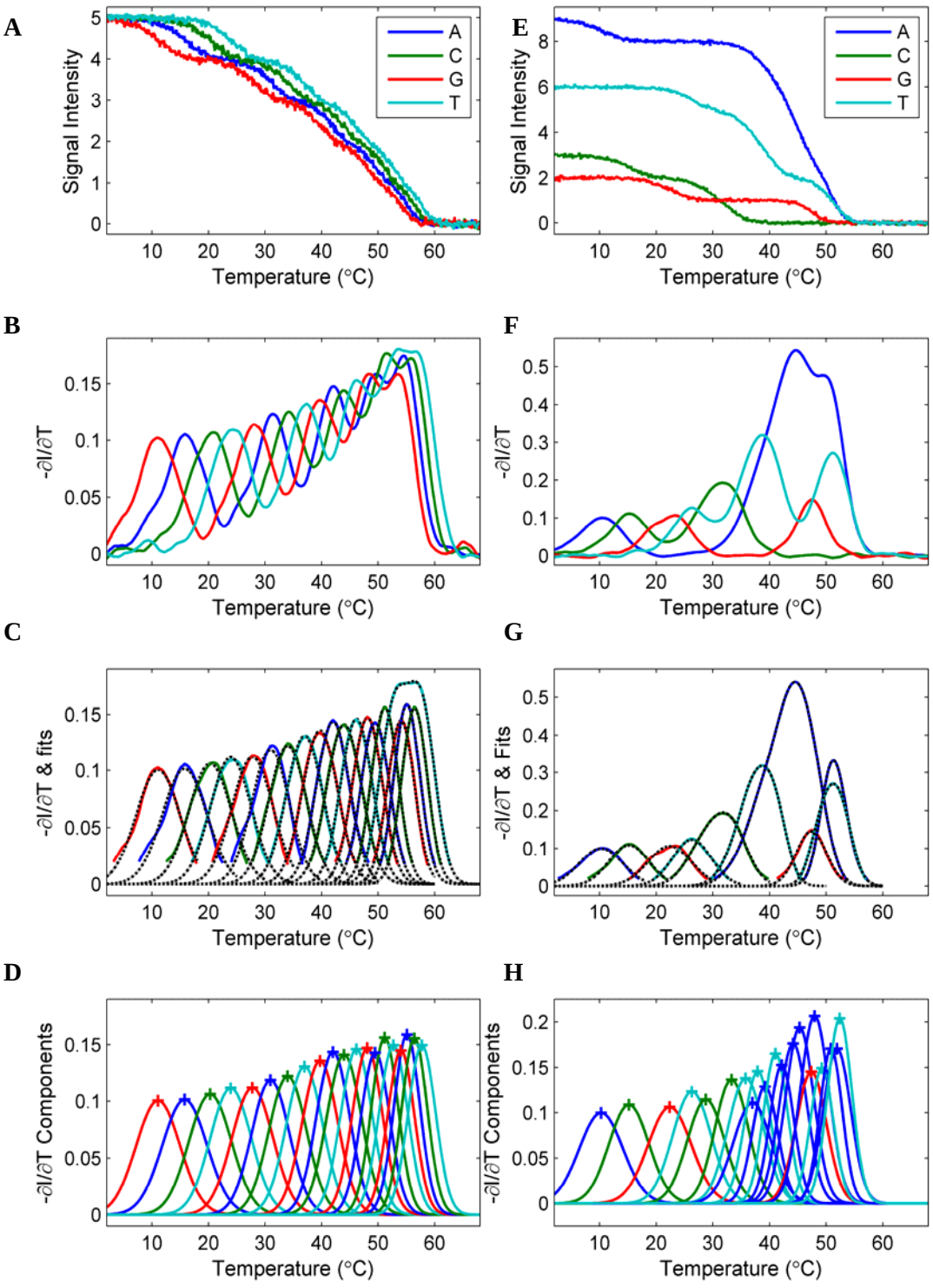
$$\sigma = -\, 4.48 \times 10^{-2}\, \mu + 5.67.$$

The initial coefficients for these parameters were set to the values determined by the above relationships at the corresponding starting point for the melting temperature. The lower and upper bounds of these coefficients were set to within 100% confidence value at the corresponding starting point for the mean position.

(4) Subtract the interference from the neighboring peaks and refine the fit. In some cases, neighboring peaks overlap. For each peak, the contributions of the neighboring peaks were subtracted by the fitted curves of all the other peaks. Then, the corrected peak was fit again to the sum of Gaussian curves with the parameters determined using the same method as in step 3. The refined fit gives a more accurate presentation of the melting temperatures of each component because the interference from the neighboring peaks was eliminated.

(5) Sort the component peaks for base-calling. Finally, the base sequence was called by sorting the melting temperatures of all of the components determined as the coefficients in the Gaussian fit. As we read from lowest melting temperature to the highest, the base sequence is called from the corresponding fluorescent channel.

Figure 6. Base-calling process for SBD. This figure illustrates the base-calling process with two examples. **(A-D)** A simple case is shown with the sequence "GTCAGTCAGTCAGTCAGTCA", where the 4 different base types are distributed evenly along the sequence. All the bases are called correctly. **(E-H)** Another case is shown with the sequence "ACGTCCTATTAAAAGATAAT". There are extensive overlaps between some of the curves. The called sequence is "ACGTCCTATATAAA GATAAT". There is a pair of substitution errors in the call. **(A) & (E)** The simulated fluorescent signal is the sum of all the contributing melting curves in the corresponding channel. A 5% random Gaussian noise is added. **(B) & (F)** The smoothed negative derivatives of the curves. Some peaks are the combination of multiple melting curves. **(C) & (G)** Each peak is fit to a sum of Gaussian curves to deconvolve the components. These figures show the results from the correction fit where interference from neighboring curves have been subtracted. The black dashed lines show that each fitted curve overlaps with its colored solid original curve well. **(D) & (H)** The components from the fit. The peaks are labeled with a cross (+) for visualization. By reading from lower to higher melting temperature, the base sequence can be determined. Blue: A. Green: C. Red: G. Cyan: T. See text for more detailed description of the steps involved in the algorithm.


A
Signal Intensity
Temperature (°C)
A
C
G
T
E
Signal Intensity
Temperature (°C)
A
C
G
T
B
-∂I/∂T
Temperature (°C)
F
-∂I/∂T
Temperature (°C)
C
-∂I/∂T & fits
Temperature (°C)
G
-∂I/∂T & Fits
Temperature (°C)
D
-∂I/∂T Components
Temperature (°C)
H
-∂I/∂T Components
Temperature (°C)

Simulations were performed over 1000 randomized sequences to evaluate the feasibility of SBD. First the SBD signal was simulated for each sequence. Then the base-calling algorithm was used to find the base sequence. For illustrative purpose, two examples and the base-calling procedure are shown. Figure 6A & E show the fluorescent intensity signals, which are the sums of all Sanger fragments labeled in the corresponding channel. A 5% Gaussian noise was added in this case to simulate measurement error. This demonstrates the expected signal from SBD measurements. The signal was then smoothed and the negative derivative was determined and smoothed as shown in Figure 6B & F. Because some peaks overlap and combine into broader peaks, the original components in each wide peak were determined by fitting them to a sum of Gaussian curves. A second fit was performed after subtracting the contributions from the neighboring peaks determined from the initial fit. Figure 6C & G show the final fit results. Each corrected peak is plotted as a colored line. As compared to the one in Figure 6B & F, the line now extends to the base of the curve since the interference from neighboring peaks has been subtracted out. The fit to each peak is shown as a black dashed line. All fits overlay well with the corrected peaks. This indicates that the fit presents the data well.

From the parameters of the fits, the Gaussian components of all peaks determined are plotted in Figure 6D & H. The peaks are marked with crosses to indicate the melting temperatures of the Sanger fragments. These curves mimic the electropherograms generated by traditional Sanger sequencing. The sequence was decoded by sorting these peaks from lower to higher temperature. An ideal case where

the 4 different base types are spaced evenly along the sequence is shown in Figure 6D. As can be seen, the height of the component peaks increases gradually as the melting temperature increases. In this case, all the peaks are well resolved and all the bases are called correctly. Figure 6H shows another case where there are extensive overlaps between some of the profiles. It is more difficult to resolve all the peaks. In this particular case, two call errors were made with the algorithm. Some of these cases could be better resolved by improving the separations between the neighboring curves. Experimentally, this can be achieved by using a combination of salt and temperature gradients.

After calling every sequence from the 1000 test sequences, the error rate was determined by dividing the number of errors by the number of bases called. Figure 7 shows the cumulative average error rate versus read length for SBD under 0.1ºC sampling frequency, and 0% and 5% Gaussian noise levels. The error rate is about 4% with a read length of 20 bases. As expected, this error rate increases linearly with read length. However, added simulated noise level has very little effect on the error rate. This indicates that the base-calling algorithm is robust against the noise that will be present in experimental data.

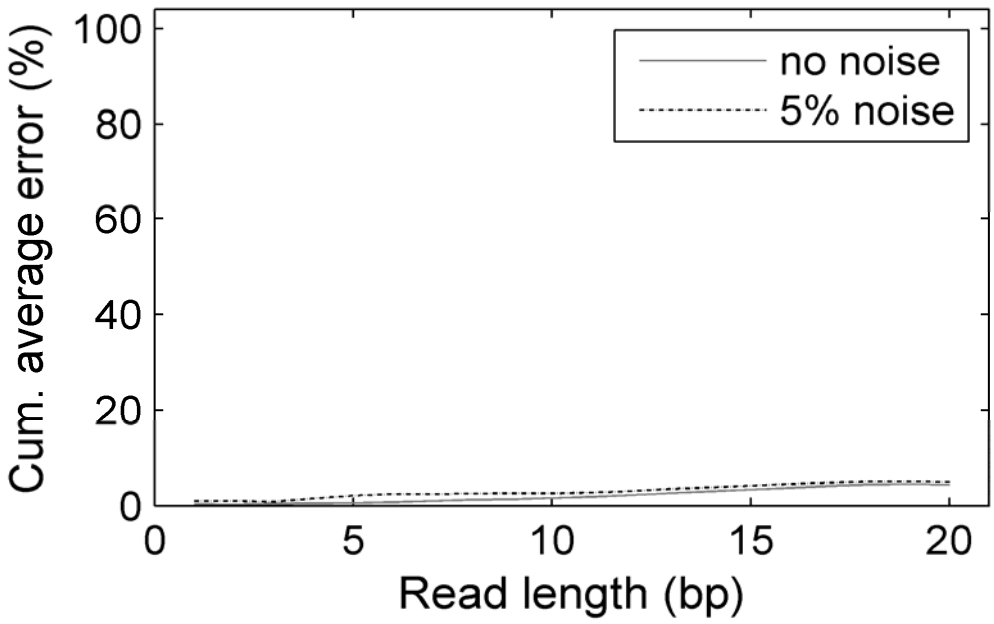


Figure 7. Error rate of SBD. The cumulative average error rate is plotted versus read length under two different noise levels with 0.1ºC sampling frequency. The error rate increases linearly with read length. With a read length of 20 the error rate is 4%. The increase in error rate as a result of 5% added simulation noise is not significant, indicating that SBD is robust against small measurement errors. Solid gray line: 0% noise. Dotted black line: 5% noise.

2.7 Summary

In this section, we have established the theoretical basis for SBD. The denaturation profiles of the DNA fragments generated by fluorescently-labeled dideoxyribonucleotides were simulated by melting curve analysis using thermodynamic principles and data. Melting curves and their negative first derivatives were plotted to show the melting temperatures as the peaks of the negative first derivatives. Both simulation and experimental results show that melting temperatures of oligonucleotides increase monotonically as each additional base is added. We have

shown how this property can be used to determine the sequence. An algorithm for base-calling has been developed to decode the DNA sequence from the intensity data. The cumulative average error rates versus read lengths were estimated with different simulated noise levels and sampling rates. Within experimentally achievable sampling frequencies, the method is robust against noise. We have demonstrated that different salt concentrations can be applied to modulate the melting curves of the DNA fragments. This allows us to have a greater control over the denaturation temperature to achieve higher accuracy and longer read lengths. This demonstrated the feasibility of SBD with simulations.

Some of the materials presented in this chapter are rearrangements from the publication: "DNA sequencing by denaturation: Principle and thermodynamic simulations", Chen, Ying-Ja and Huang, Xiaohua, *Analytical Biochemistry*, 384(1): 170-179 (2009). The dissertation author was the primary investigator and author of this paper.

Chapter 3 A High-Speed Fluorescence Imaging System with Integrated Fluidics and Temperature Control

3.1 Introduction

In this chapter, I describe the construction of a fluorescence imaging system capable of high-speed scanning that is integrated with a biochemical reaction chamber. All components except a custom-made reaction chamber are commercially available parts. The system has many potential applications in high-throughput technologies. Fluorescence detection can be multiplexed to many different fluorescence channels. The availability of many fluorescent molecules that can be conjugated to various macromolecules allows the system to be widely used in the detection or tracking of biological molecules and processes. The high-speed fluorescence scanning system allows the screening of many molecules in a short period of time. This can be used as biosensors for molecular diagnostics, genotyping, or other genomics technologies. In the following chapter, we will show how it is utilized for a massive parallel DNA sequencing application.

3.2 Fluorescence imaging system

Our fluorescence imaging system is based on an inverted fluorescent microscope with motorized configuration (Zeiss Axiovert 200M). Several accessories were chosen for optimal high-speed scanning, including a high-speed 5-channel light switching device (Sutter Instruments DG-5), a linear-encoder motorized stage (LUDL BioPrecision 2), the motorized z-drive internal to the Axiovert 200M, and an EMCCD

camera (Andor iXon^{+}). A quadband-pass filter set (Semrock FF01-440/521/607/700-25) was used for four color imaging without having to change the reflector turret or any excitation or emission filter wheels. The fluorescent channels were chosen by the excitation wavelength generated by the DG-5 light switching device.

The optics on the microscope were optimized for sensitivity and speed for detection of fluorescent molecules on a surface. With more sensitive optics, the exposure time on the camera can be reduced, thereby achieving a faster imaging speed. Plan-apochromat objectives were chosen for their high numerical apertures (NA) and correction for aberrations. The magnification and NA of these objectives are 10X/0.45NA, 20X/0.8NA, 63X/1.4NA, and 100X/1.4NA. The lower magnification objectives (10X and 20X) were used for high-speed scanning applications when there were sufficient fluorescent molecules for detection. This allowed a larger field of view to be acquired, minimizing the number of fields, which reduced the time spent on scanning. This results in the reduction of the number of the stage movements. The higher magnification objectives were used for high-sensitivity applications especially when single-molecule detection is desired. The Andor EMCCD camera is ideal for these applications because the fluorescent intensity signal can be amplified through the linear electron-multiplying (EM) gain to utilize the maximum dynamic range of the 14-bit camera. For high-speed applications, frame-transfer mode was used so that the maximum speed of the 35MHz camera was fully utilized.

In this imaging system, high sensitivity is desired not only to minimize the exposure time for image acquisition, but also to minimize the number of fluorescent

molecules required for the detection so that the samples can be prepared with ease. An EMCCD camera amplifies the signal through EM gain so that signal below the limit of a normal camera can be detected. Maximum signal-to-noise ratio (SNR) can be reached by adjusting the EM gain to the proper level. By applying EM gain, single fluorescent molecules can be detected through a 63X objective in about one second to achieve at least an SNR of 2. Using the same objective lens, 100 molecules can be detected in 100 ms with an SNR of 10. When using a 20X/0.8NA objective, several hundreds of milliseconds are required for the detection of 100 molecules with an SNR of 2 to 5. When using the 20X or 10X objectives, it is more desirable to prepare 10,000 molecules or more to utilize the full dynamic range of the camera while keeping the exposure time within 100 ms for higher speed imaging. These properties provide a guideline for the preparation of samples to be imaged. With the combination of high NA objectives and an EMCCD camera, the desired SNR can be reached by imaging on a reasonable number of fluorescent molecules without compromising the imaging speed by using a high exposure time.

High-speed scanning has been implemented with the parameters described below. Generally, four images are taken in the four fluorescent channels with 38~100 ms of exposure time each, depending on the signal intensities of the fluorescent molecules. The time to switch between fluorescent channels using the DG-5 is 1 ms and the time to move the stage from one position to another neighboring position is 50~100 ms. The time to image the entire surface of a fluidic channel on the device depends on the number of fields to scan through. With the 10X objective and the 8 μm

pixel size on the 1004 × 1002 pixel camera, each field of view is 803.2 × 801.6 μm^2. Then there is a total of 546 fields to scan through for the 5 × 62 mm^2 channel used or 1875 fields to scan through the entire 20 × 60 mm^2 flow cell. If four colors are used, the total time to scan through one fluidic channel is 164~273 seconds, which is within 5 minutes. The time it would take to scan through the entire flow cell would be 563~938 seconds, which is around 10~15 minutes. These simple calculations indicate the imaging throughput of this system. The speed of most fluorescence-based high-throughput detection technologies is limited by the imaging throughput. This system provides optimization over imaging speed and sensitivity for maximal throughput for many of these applications. This fluorescent imaging system was automated by a program written in C++. The program controls every part of the system including the microscope, the stage, the camera, the temperature controller, and the fluidic components.

3.3 Custom chamber design and temperature control

In order to construct a high-speed fluorescence imaging system for high-throughput DNA sequencing and other biomedical applications, a biochemical reaction chamber is needed to satisfy several requirements: 1) It has to allow fluorescent imaging and scanning. 2) Reagents and wash buffers have to be delivered to the reaction chamber through fluidics design. 3) There must be active temperature control to reach and maintain the optimal reaction temperatures during each step of the sequencing process. In order to satisfy all of these requirements, we have designed a custom-made reaction chamber.

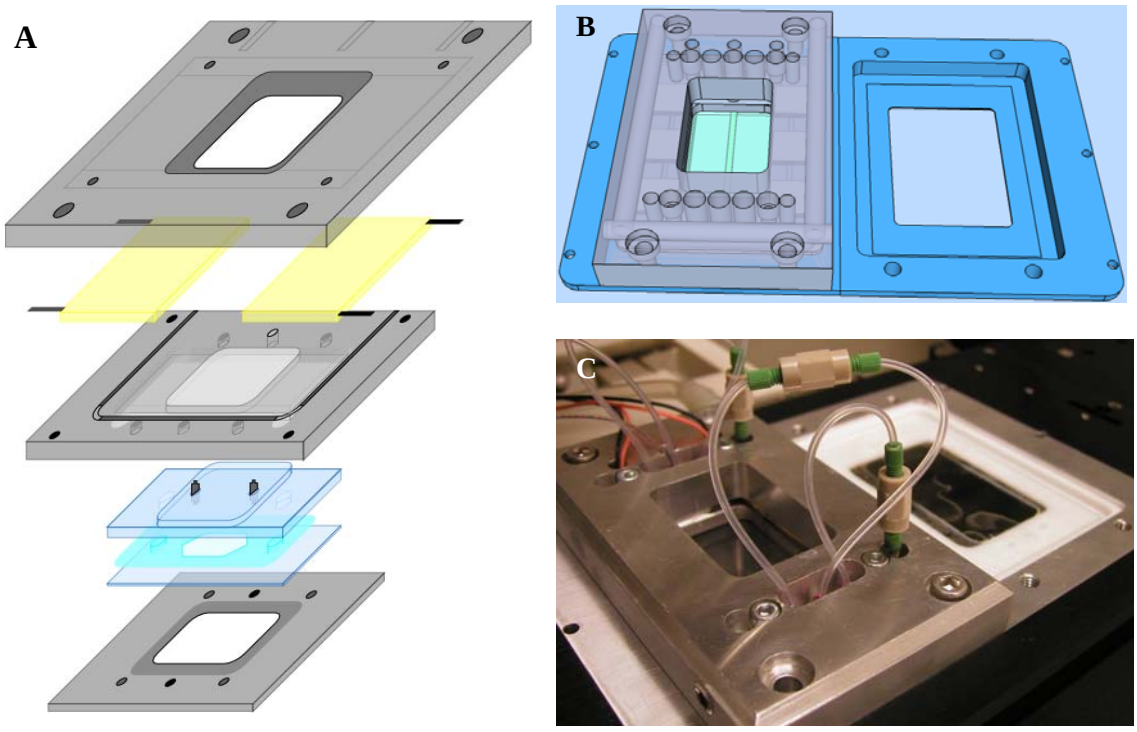


Figure 8. Chamber Design. (A) A schematic of the initial chamber design that illustrates each component of the chamber. In blue is the double-sided tape channel sandwiched between the cover slip for imaging from the bottom and a top viewing slide. These pieces are held together by the metal pieces on the top and bottom which fits on top of the microscope stage. The thermoelectric modules for temperature control are shown in yellow. The metal piece on top of that is the heat sink. **(B)** A three-dimensional view of the final chamber design in the assembled form. The design in this version differs from the initial one by some mechanical dimensions and an insulating piece underneath the cover slip. **(C)** A picture of the chamber.

Figure 8 illustrates the design of the custom-made chamber. Figure 8A shows a schematic drawing of each component in an initial chamber design. The reaction surface is a cover glass derivatized with functional groups to capture the beads or sequencing templates. The reaction chamber was created by sandwiching a double-sided tape (Scapa 702, 116.84 μm thick) between the cover glass and a glass slide with

the shape of the channels cut out by a CraftROBO cutter. There are holes on each side of the channel on the glass slide allowing reagents to flow through the chamber. The reaction chamber was set on an aluminum stage adaptor allowing it to fit on the stage of an inverted microscope for fluorescent imaging. On the top of the reaction chamber is a thin aluminum piece that interfaces it with the thermoelectric (TE) modules. Heating and cooling are controlled through these TE modules. Another larger aluminum piece covers the TE module and functions as the heat sink. The center of the heat sink was cut away for observation of the reaction chamber during the imaging process by eyes. This piece was screwed onto the stage adaptor. Water is flown through this piece by an external circulator (Julabo F25-HE) which keeps it at constant temperature. The screws used to hold this piece together with the TE module-interfacing metal was insulated with a PEEK screw to separate the hot surface from the cold surface. The insulation between the reaction chamber and the stage adaptor was achieved by a Teflon layer in between the two pieces. In order to bring the temperature in the reaction chamber to the desired temperature during washing, the fluid lines are connected to the embedded stainless steel tubings that circulate around the TE module-interfacing piece before entering the reaction chamber. Upchurch 062" tubing connectors were used to connect soft tubings to these metal tubings while Lee Minstac connectors were used to connect the soft tubings to the TE module-interfacing metal, which then directs the fluidics into the reaction chamber with an o-ring as sealer.

The final design of the chamber as shown in Figure 8B & C is mostly the same as the initial design described above with a few improvements. The final chamber was

designed to be more compact so that two chambers can be fit onto one system. When one chamber is being imaged, the sequencing reaction can be performed in the other chamber to fully utilize the imaging capability of the system. A two inch by three inch cover glass was used as the reaction surface so that a larger imaging window can be used to maximize throughput. Eight compact TE modules (Marlow MI1023T-02) were used to provide the necessary heating and cooling power. Three channels were used in each chamber. Each of them is 5 mm × 62 mm. The size and shape of these channels could be changed easily by altering the drawings for the CraftROBO cutter which cuts the double-sided tape. This final chamber design allows high-throughput sequencing reactions to be performed in one compact chamber.

The temperature control of the system was accomplished by setting TE modules on the reaction chamber interfaced through a metal piece. An aluminum piece was chosen because it is easier to machine. However, other metals such as stainless steel, which have better thermal transfer properties, may also be used. The thickness of this piece was minimized to ensure efficient transfer of heat to the chamber while minimizing the power required by the TE modules. A total of eight TE modules (Marlow MI1023T-02) were packed on the device with the hot side facing the chamber and the cold side facing the heat sink because heating of the chamber is desired in most applications. Each pair of them was connected in series and then the four pairs were connected in parallel. This configuration provided the large output power necessary without exceeding the maximum voltage of the TE modules as they were connected to a 13.8V, 260W power supply (RadioShack). The output was

controlled by our program through a temperature controller (TE Technology TC-24-25). The temperature controller can maintain the temperature of the chamber at a certain temperature using proportional-integral-derivative (PID) control or provide a certain output voltage to the TE modules. A thermistor was fixed to the aluminum chamber piece to record the temperature in the chamber for feedback control. It was connected to the temperature controller. Different temperature ramping profiles can be achieved by programming the device from a computer.

For SBD, a linearly increasing temperature profile is desirable for the measurement of the denaturation profiles of DNA fragments. Several output profiles have been tested and a linear output voltage was chosen. Figure 9 shows how the temperature profile was controlled over time. A linear increase in the negative output power would generate the desired linear temperature profile. As shown in Figure 9A, the output power was controlled to increase linearly through time with different initial powers and slope.

Figure 9B shows the corresponding temperature profiles. In each case, the temperature increases in a nearly linear profile. When the slope parameter for the output voltage was fixed, the temperature profile with the defined speeds or slope can be achieved. This can be seen in the blue and light green lines, and the red and magenta lines. Because different initial voltages were used, the temperature profile entered the linear phase at a different time, but increased at the same speed. Although most experiments began at the same temperature of 37ºC, if the initial voltage was too low, the temperature remained constant or decreased initially before the linear phase

began. This is observed in the dark green and cyan lines. If the voltage reaches 100% before the target temperature was reached, the temperature would fail to maintain its linear profile and increase at a slower rate or remain constant. This was observed in the black and magenta lines. Each of these temperature profiles was fit to a straight line. The R^2's of fit were 0.998 indicating that the temperature profiles were very linear. In Figure 9C, the dependence between the slope parameter for controlling the output voltage and the slope of the temperature increase was shown. There is a linear relationship:

$$\text{(slope of temperature increase)} = -0.044\ \text{(slope parameter)} + 0.0026.$$

These results indicate that a linear temperature ramping profile can be achieved with the desired rate by varying the slope parameter and initial output voltage. The time for the temperature ramp to reach a certain target temperature can also be determined by the plots in Figure 9.

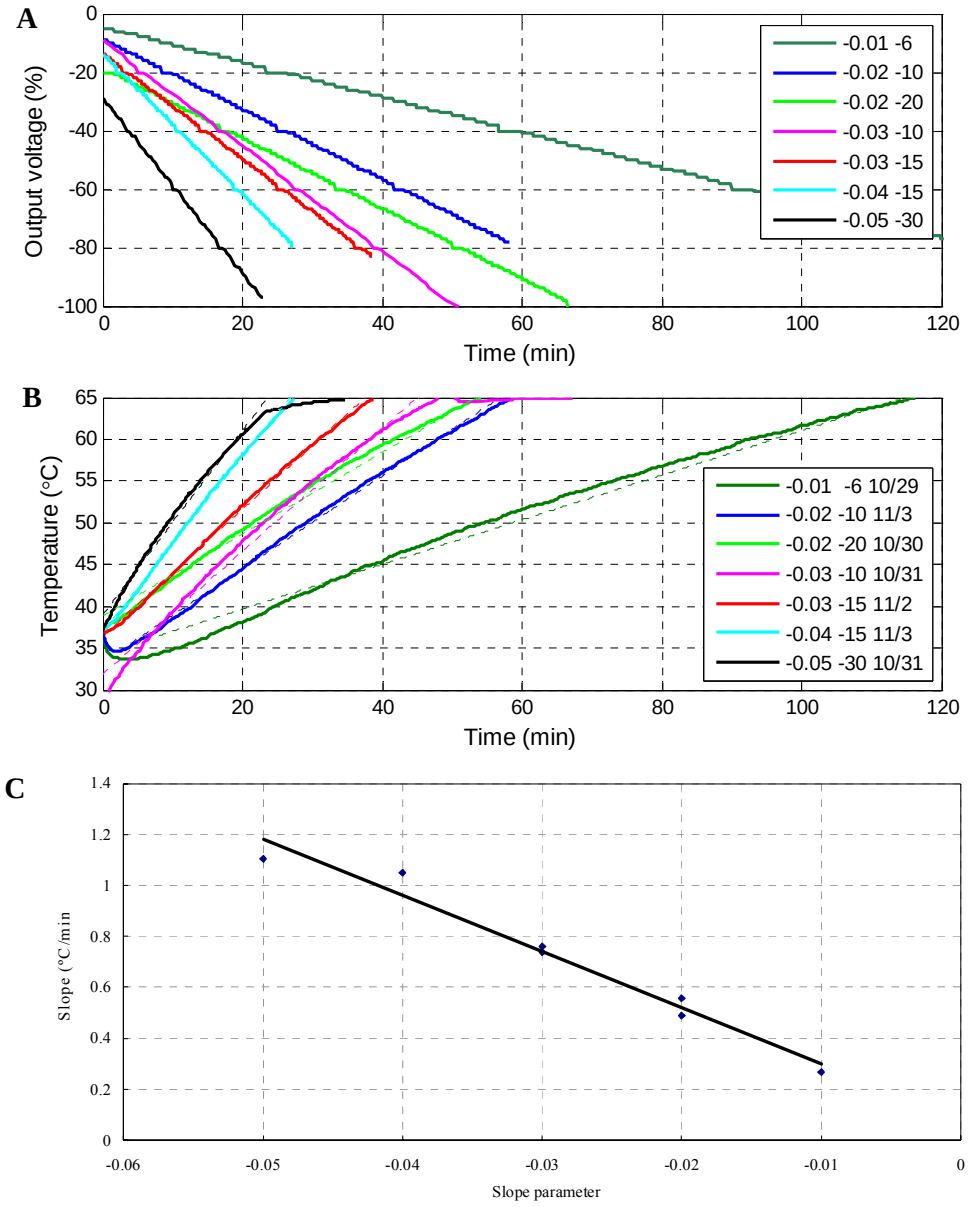


Figure 9. Temperature profile can be controlled precisely. (A) The negative output voltage was increased linearly with an arbitrary slope parameter varying from -0.01 to -0.05 and starting voltages of -6% to -30% of the total 14 V. **(B)** The resulting temperature profiles corresponding to the linear output voltages are close to being linear. The larger the slope parameter, the faster the temperature increases. Solid lines: temperature profiles. Dotted line: the fit of temperature profiles to straight lines. **(C)** The slope parameter used in increasing output voltage determines the speed of linear temperature increase.

3.4 Fluidic design

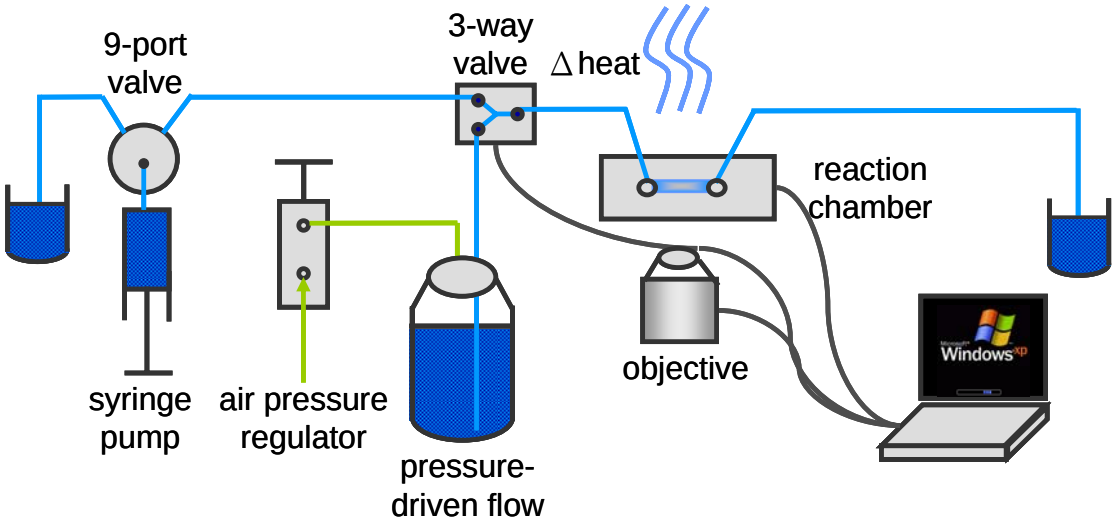


Figure 10. Schematic of the fluidic system. The syringe pump and the nine-port valve were used to deliver reagents into the reaction chamber through the choice of a three-way solenoid valve. The other choice from the three-way valve was a passive flow of wash buffer driven by a regulated constant air pressure. The wash buffer was contained in a bottle with larger volume which provided a constant flow throughout the entire experiment without generating accelerations and decelerations that would be generated by a syringe pump.

The fluidic design of the system is shown in the schematic diagram in Figure 10. A syringe pump and a nine-port valve (Cavro XL3000 and Tecan Pump XR 9+) was used to direct 9 different reagents and/or wash buffers to the reaction chamber as programmed. For simplicity, only one reagent is shown to be connected in this diagram. A solenoid valve was used to choose between the delivery of reagents and wash buffer through the syringe pump or a constant fluid flow that could remain continuous through a long period of time until the whole liter of buffer was depleted. This constant fluid flow was driven by air pressure which was maintained constant

through a nickel-plated brass subminiature air-pressure regulator (McMaster-Carr 3834T51). After the reaction, the solutions were collected into a waste bottle. The selection of reagents and the washing methods were controlled by the same program used for the fluorescence imaging system and the temperature controller.

3.5 Integration and performance

When integrating a system with imaging, temperature, and fluidics, several issues arise and must be solved. This includes the elimination of bubbles formed in the fluid line or reaction chamber caused by heating, the change in fluorescence intrinsic to fluorescent molecules due to temperature change or photobleaching effect, and the change in focus position due to temperature change. We have characterized each of these phenomena and calibrated the system so that each of them was corrected for our application on high-throughput DNA sequencing.

3.5.1 Bubbles in fluid lines at high temperatures

In SBD, the denaturation profiles of the Sanger reaction products are measured by detecting the fluorescent intensity while the temperature is increased gradually. When the temperature increases, the mini air bubbles in the fluid line expand and become visible. These bubbles not only reduce the volume for reagents to be delivered to the reaction surface, but also alter the imaging properties during fluorescent detection. Moreover, as bubbles expand, their size can possibly grow to the extent that the entire chamber is covered by air rather than buffer. In order to eliminate these bubbles, the fluid lines must be primed at

higher flow rate to break up bubbles and push them away. A surfactant, 0.01% of Trion X-100, was added to the wash buffer to facilitate this process. In addition, constant fluid flow during the ramp of temperature ensured that bubbles were washed away constantly during the process.

3.5.2 Photobleaching effect

In SBD, images are taken at the same position and fluorescent molecules multiple times during the temperature ramp. When excited multiple times, organic fluorescent molecules may undergo a process termed photobleaching, where thermal energy conversion takes place causing the loss of fluorescence. The photobleaching effect results in an exponential decay in fluorescent signal through time. The measurements of the photobleaching effect for several fluorescent molecules are shown in Figure 11. In these experiments, fluorescently-labeled molecules were spotted and immobilized on the imaging surface. An image was acquired at the edge of the spot every second with 100 ms of exposure time. The fluorescent intensity was averaged over a rectangular region in the spotted area. In each subsequent image, the fluorescent intensity of the same region was averaged and normalized to the intensity of the first image. In Figure 11A, the photobleaching curve of Alexa 546 was measured. The fluorescent intensity dropped by 50% after 250 images, which corresponds to 25 seconds of exposure. In Figure 11B, the photobleaching curve of Alexa 647 was measured. These images were taken with 300 ms of exposure time. The fluorescent intensity dropped by 50% after about 30 images, which corresponds

to 9 seconds of total exposure time. These measurements indicate that using 38~100 ms of exposure time, Alexa dyes can be imaged a couple hundred times before most of the fluorescence is photobleached. This is important and reasonable for SBD because a couple hundred sample points are enough to represent the denaturation profile of short DNA fragments. However, when analyzing the data, the photobleaching effect must be subtracted from the signals to obtain the true denaturation profiles.

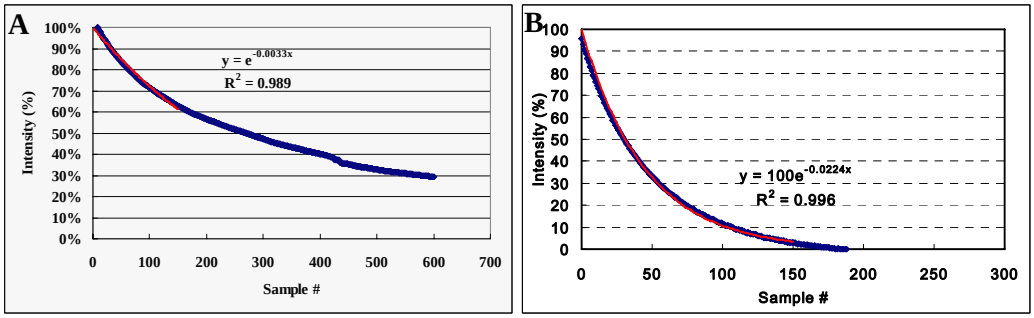


Figure 11. Photobleaching of fluorescent molecules. Photobleaching effect of **(A)** Alexa 546 and **(B)** Alexa 647. The fluorescent signals from Alexa dyes decrease exponentially when they are exposed for many times. The time that the fluorescent signals drop by 50% is at 100~300 exposures which corresponds to 9~25 seconds.

3.5.3 Temperature effect of fluorescent molecules

Another issue that arose during the integration of fluorescent imaging and temperature ramping was the intrinsic change in fluorescent quantum yield at different temperatures. When the temperature increases, the fluorescent quantum yield of the fluorescent molecules decreases. We have measured this effect by monitoring the fluorescent intensity from fluorescent molecules in a cuvette with

a fluorometer (PerkinElmer) while changing the temperature. The temperature was controlled by changing the temperature of the circulating water that flows around the cuvette. One milliliter of fluorescently-labeled oligonucleotide solution was prepared and dispensed into a cuvette with a 1-cm light path. Because the volume being excited during measurement was quite small compared to the total volume of fluorescently-labeled oligonucleotide solution, we reason that photobleaching effect had minimal contribution to the decrease in fluorescence. Therefore, the decrease in fluorescence shown in Figure 12 below was mostly caused by the increase in temperature. Figure 12A~D shows the decrease in intensity of organic fluorescent molecules Alexa 488, 546, 647, and 750. As shown, the fluorescent intensity decreases linearly with temperature. Notice the y-axis does not start from zero and that the fluorescence only dropped by about 15% from 20 to 55ºC. Figure 12E shows the temperature effect on fluorescence for 20 nm nanoparticles containing the equivalent of 180 fluorescein molecules. Unlike the temperature effect curves of organic molecules, that of the nanoparticles increases exponentially. We believe this was due to the thermal expansion of the nanospheres so that the spatial constraints of the fluorescent molecules were loosened, allowing the fluorescence to be more efficiently excited and emitted at a higher temperature. These nanoparticles may be useful for SBD experiments because they may provide higher signal intensity at higher temperature.

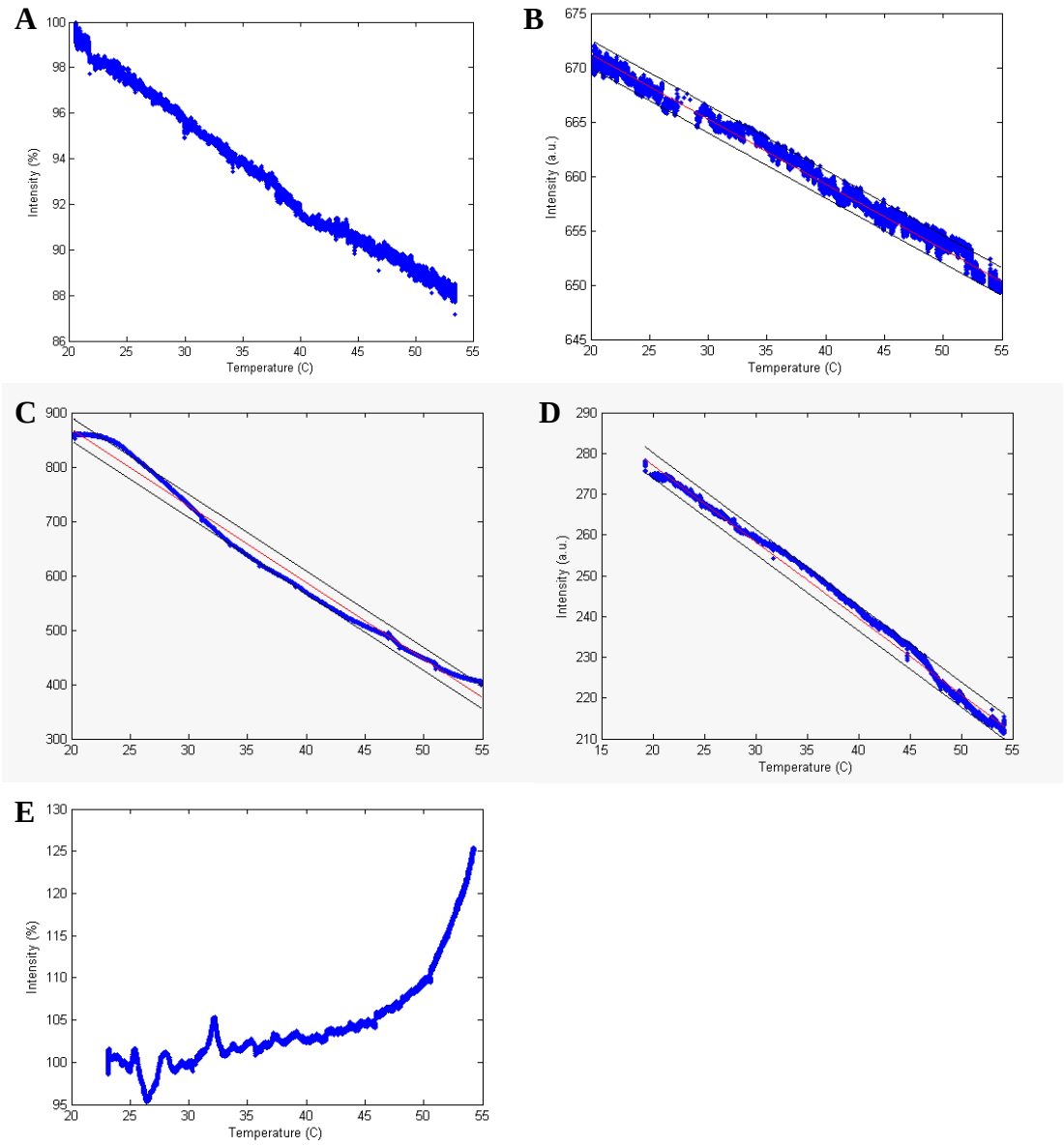


Figure 12. Temperature affects fluorescence. The change in fluorescent intensity versus temperature of **(A)** Alexa 488, **(B)** Alexa 546, **(C)** Alexa 647, **(D)** Alexa 750, and **(E)** 20 nm nanoparticles (Invitrogen FluoSphere). The red lines are the fits to linear lines. The black lines are the linear lines plus and minus two times the standard deviation. The fluorescent intensity of Alexa dyes decrease linearly as temperature increases. Note the y-axis does not start from zero. From 25 to 55ºC, the fluorescence has only dropped by 15%. For nanoparticles with embedded fluorescent molecules, the fluorescent intensity increases as temperature increases.

3.5.4 Autofocus

When acquiring images over a period of time, the focus position may drift away. This issue is even more serious if the images are acquired as the temperature in the chamber increases because thermal expansion will cause the chamber to expand and change the z-position of the imaging surface. The drift in focus position as the temperature increase is shown in Figure 13. As can be seen, the focus position decreases linearly as the temperature increases. This linear relationship was used to adjust for the focus when the temperature was changed during the experiment. In each experiment, the focus position was first found manually. Then, the focus position was adjusted according to the initial focus position and difference between the new temperature and the initial temperature.

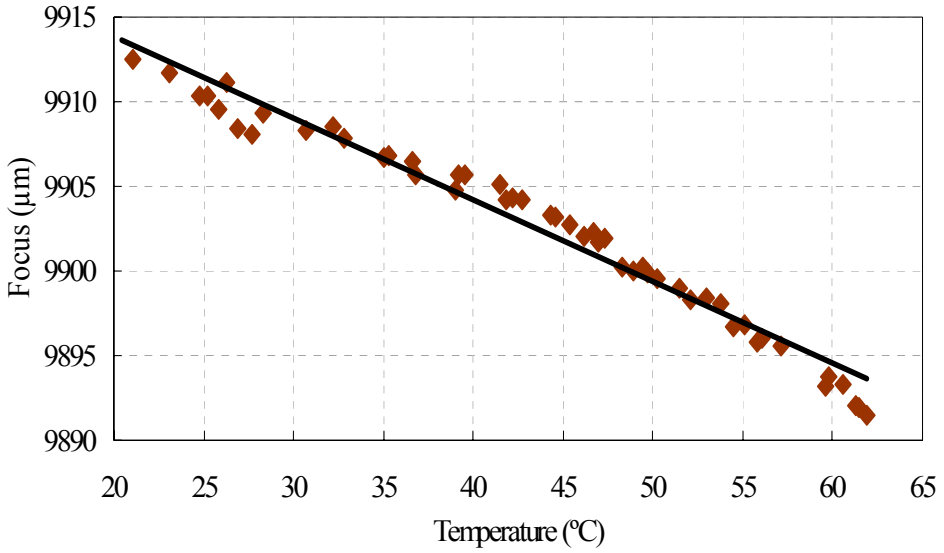


Figure 13. Focus position changes as temperature increases. Due to thermal expansion of the system, the focus position decreases linearly as temperature increases.

Although the focus position varies linearly with temperature, the focus position can be slightly different from that predicted by the linear relationship. Moreover, finding the correct focus plane is crucial for any fluorescent application. It is especially sensitive when the numerical aperture of the objective is high and the focal depth is narrow, which is the case for most high-sensitivity applications like SBD. In these cases, it is important that the correct focus position is found with sub-micrometer accuracy.

We have developed a method to find the focus position through an automated routine. First, the objective was moved up with a set velocity. During the constant velocity period, several images were taken. Then, the auto-correlation of each image was calculated. The image with the largest auto-correlation value was assumed to be the image with maximal contrast taken at the focus position. Finally, the z stage was moved to the position where the image with the largest auto-correlation value was taken.

In this autofocus process, there were several parameters customized to each imaging condition including the range to find focus, the center of the range, the step size between each image taken, and the imaging parameters for taking the pictures. When finding the initial focus, a coarse autofocus routine was performed with a larger range and step size followed by a fine autofocus routine with finer range and step size around the focus position found in the coarse focusing process. In general, the step size of the fine autofocus routine was chosen to be less than the focal depth of the objective lens and the range was chosen to span less than 20

images. The former ensures that the correct focus position will be found and the latter minimizes the time required to find the focus.

In the case when focus has drifted due to temperature, the center of the autofocus range is chosen according to the initial focus position and the change in temperature. If the temperature has not changed by more than a few degrees Celsius, then the range for autofocus would not be larger than a few micrometers, which required less than 10 images to be taken. Then the autofocus routine would not take longer than a few seconds during scanning.

When scanning through a large area on the surface, the focus position would change from one end of the surface to the other. Three autofocus methods have been implemented to adjust for this situation under different circumstances. When the image region was small, autofocus was performed at one spot adjacent to the imaging region. Because the positions were adjacent to each other, the focus position at the imaging region remains the same as the focusing spot. Autofocusing at the nearby spot prevented the imaging region from being photobleached by the autofocus routine while finding the correct focus for image acquisition. When the imaging region was slightly larger, autofocus was performed along the perimeter of the imaging region. In order to interpolate for a given imaging spot, the interpolation splines of the focus position was calculated using first the two vectors of the autofocusing position in the x-direction and then the y-direction. The focus position at each imaging spot was approximated by the average of the two interpolations calculated. When the imaging region was large,

autofocus was performed in a grid across the imaging region. The focus at each imaging spot was interpolated by finding the two-dimensional spline. These autofocus processes were performed before each round of image acquisition took place during an experiment.

3.6 Summary

With the integration of fluorescent imaging, temperature control, and fluidics, many high-throughput biological assays could be performed on this high-speed imaging system. Several issues in the integration of these components have been solved. This system can be customized to perform biochemical reactions under constant or varying temperatures with or without buffer washing while being monitored through fluorescent imaging. The proper autofocusing and scanning routines could be chosen for each application tailored to the reaction conditions.

Chapter 4 Experimental Proof of Concept for Sequencing by Denaturation

4.1 Measurement of denaturation curves in solution

In order to establish the proof of concept for SBD, we first measured the denaturation curves of 8 oligonucleotides in solution. The melting curves of the oligonucleotide probes were measured with a UV-Vis spectrometer (PerkinElmer Lambda-20) by measuring the absorbance at 260 nm through time while the temperature is gradually increased. The samples were placed in a cuvette with a flow cell formed by the thin walls around the sides of the cuvette. The water temperature in the flow cell was controlled to within ±0.1 ºC using a Julabo F25-HE circulator with an external temperature probe in the cuvette. The denaturation curves of eight oligonucleotides consist of the first 20 bases and addition 1 to 8 bases of the sequence "CCATCAGTCATGTACG AAGTCAGTCATG" were measured. The samples were prepared by combining 650 nM of each probe with 650 nM of a common template sequence "TAGCATGACTGACTTCGTA CATGACTGATGGTCGA" in a 33 mM phosphate buffer, pH 7.2, which has an equivalent of 49 mM of monovalent cation concentration.

In order to mimic the denaturation profile of the Sanger sequencing products, the oligonucleotide probes that end in the same base type were combined to obtain the SBD signals from an 8-base read: the 22mer and 26mer for A, the 21mer and 25mer for C, the 23mer and 28mer for G, and the 24mer and 27mer for T. In each solution,

the concentration of each of the two probes was 325 nM and the common template concentration was 650 nM so that each probe could hybridize to one template.

The melting curves were fit to a sigmoid curve to determine its baseline and upper line for normalization. After normalization, the denaturation profiles that mimic the SBD signals were analyzed by the base-calling algorithm described above.

The melting curves of the 8 oligonucleotides are shown in Figure 14A. The shape of these curves overlap with the predicted melting curves very well. The melting temperatures of the oligonucleotides are shown in Figure 14B. The melting temperature of the oligonucleotides increases monotonically as the length of the oligonucleotide increases. The base-calling process is shown in Figure 14C to 6F. Figure 14C shows the denaturation profiles in 4 channels, each of which contains two component oligonucleotides. The corresponding negative derivatives of these denaturation profiles are shown in Figure 14D. The fit to a sum of Gaussian curves and the component peaks are plotted in Figure 14E & 6F. The shapes of the curves resemble those of the simulated negative derivatives. After performing the base-calling on these data using the algorithm developed in section 2.6, the original sequence of CAGTCATG can be determined correctly. These measurements confirm that the model developed previously predicts SBD data well and that the base sequence of an octamer can be determined using the base-calling algorithm.

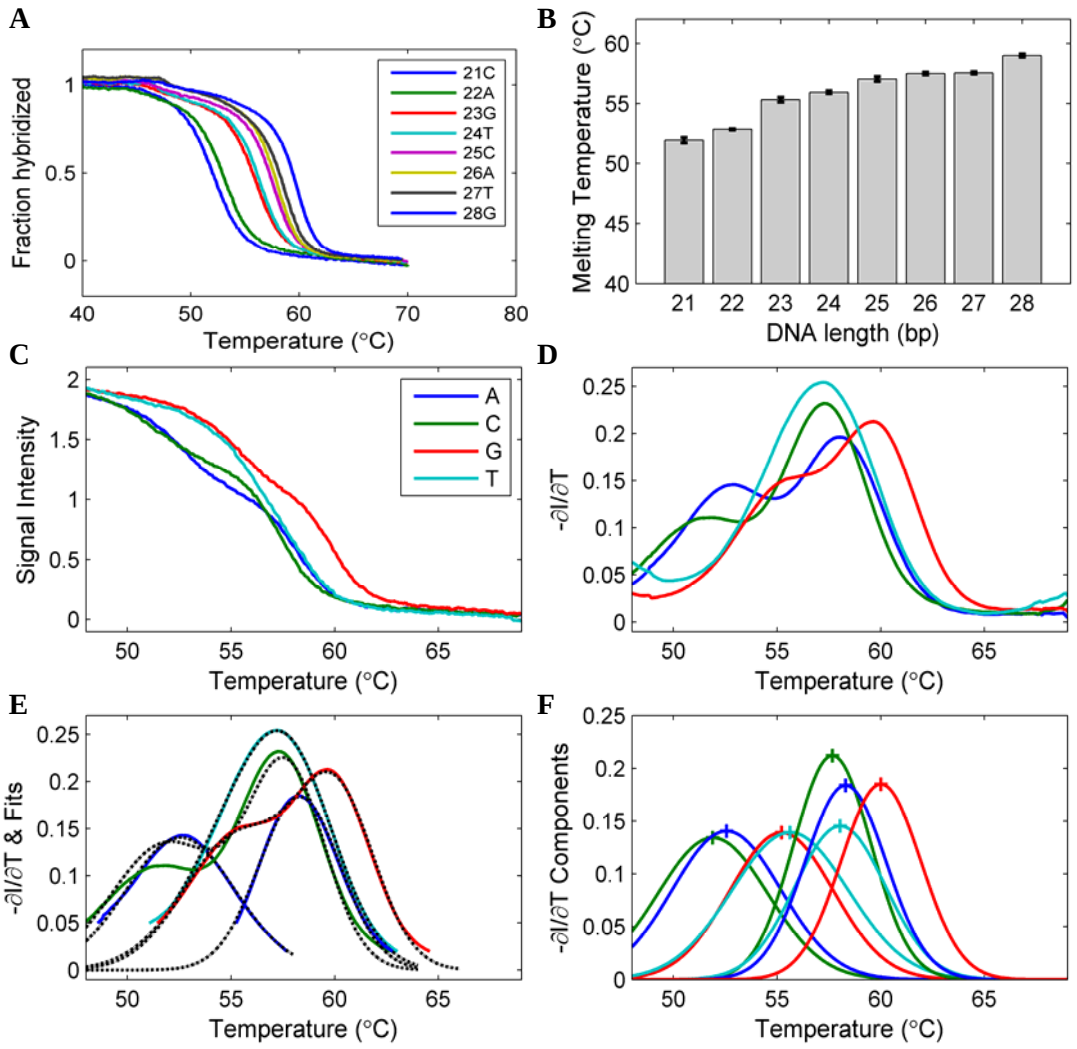


Figure 14. Experimental measurement of melting curves and SBD signal in solution. (A) The melting profiles of the 8 individual oligonucleotides with 21 to 28 bases from the sequence "CCATCAGTCATGTACGAAGTCAGTCATG". **(B)** The melting temperatures of the oligonucleotides. It is obvious that the melting temperature increases as the length of the oligonucleotide increases. **(C)** Solution measurements mimicking the SBD process. To mimic the denaturation profile of Sanger fragments, the oligonucleotide probes that end in the same base type were combined to measure the SBD signals for an 8-base read: the 22mer and 26mer for A, the 21mer and 25mer for C, the 23mer and 28mer for G, and the 24mer and 27mer for T. **(D)** The negative derivative curves of the denaturation curves. These curves were used for the base-calling process. **(E)** The fitting of the profiles with a sum of Gaussian curves. **(F)** The resolved individual components used to determine the base sequence. Blue: A. Green: C. Red: G. Cyan: T.

4.2 Introduction to the experimentation on the surface

In the previous sections, we have demonstrated the proof of principle of SBD through experimental measurements of the denaturation curves of 8 oligonucleotide probes in solution. However, performing denaturation measurements on the surface is necessary for SBD to be useful as a high-throughput sequencing method. In this section, we will demonstrate how the integrated system constructed in the previous chapter can be used for DNA sequencing, especially SBD.

Recall that SBD is performed by conducting the Sanger sequencing reaction on a surface and subsequently denaturing these DNA fragments while monitoring the denaturation profile. In our setup, the Sanger reaction can be performed with fluorescently-labeled dideoxyribonucleotides and the denaturation profiles are detected by monitoring the fluorescent signal on the surface while the temperature is gradually increased. The base sequence is determined by the drop in fluorescence in the corresponding channel as each Sanger fragment denatures and leaves the surface. For SBD to be feasible, we must be able to detect the change in fluorescent intensity when one species of Sanger fragment leaves the surface and the time or temperature difference between this event and the denaturation event of the next Sanger fragment with only one additional base. These requirements pose great challenges for the fluorescent imaging system and temperature control.

To tackle these challenges one at a time, we started by measuring the denaturation curves of a single species of oligonucleotide fragments, by monitoring the change in fluorescence of a fluorescently-labeled oligonucleotide probe hybridized

to a template immobilized on the surface. We proved that accurate measurement of denaturation profile could be performed on a surface with our imaging system. Then we showed that single base differences in the oligonucleotide sequences could be distinguished experimentally. The results are shown in section 4.4. During the measurement of the denaturation profiles on the surface, many images were acquired. To perform SBD, we need to track the fluorescent signal intensity from each of the templates in the series of images. The locations of the templates in each image must be determined, aligned, and normalized to reveal the denaturation profile. In section 4.3, we describe the image processing and data analysis methods in detail. Finally, we analyze the performance of SBD by calculating the imaging speed and throughput for this system in section 5.1.1, which demonstrates that SBD is a competitive method compared to other currently available technologies.

4.3 Image processing and data analysis

The data acquired from the measurement of the denaturation profile on the surface are a series of images. These images contain the bright spots from the templates on the microbeads which were immobilized on the surface. The denaturation signal of a species of oligonucleotide probes is the fluorescent signal intensity profile from one microbead tracked over all the images. We have written an ImageJ (NIH) software plugin using Java (Sun Microsystems) to process the images into a matrix of fluorescent intensity signal of each microbead over multiple time points. The matrix was then analyzed by a script written in MATLAB to plot the denaturation curves or other statistics of interest.

The process for extracting fluorescent intensity data from the images involves multiple steps as described below.

(1) Subtract the background. The background was subtracted for each image by a rolling ball algorithm with a rolling ball radius of 25.

(2) Threshold the image. The first image in the series was segmented to determine where the bright spots were. The threshold image was dilated with a count of two where every pixel that had two adjacent pixels being part of an object will be labeled as part of the object. The dilated threshold image was saved and used for particle analysis.

(3) Analyze particles. For every image in the series, the threshold image generated by segmenting the first image was used for particle analysis to determine the pixel location of each microbead and its signal intensity. The sizes of the particles recorded were restricted to 16~24 or 16~72 pixels depending on the imaging parameters. This criterion excluded bright dust particles, microbead aggregates, and those on the edges of the images.

(4) Register the images. Because the images may be shifted by several pixels from the first image during the experiment, each image was first registered to the first image before analyzing the particles. The registration was done by calculating the cross-correlation of the current image with the first image and finding how many pixels away the largest cross-correlation value was from the center pixel. The image was then shifted by that many pixels to align with the first image and the

threshold image. Because the threshold image has been dilated, if a microbead did not occupy the exact same position as in the first image, the pixels covering the entire microbead will still be captured.

(5) Calculate the signals. The fluorescent intensity of each microbead was calculated by the integrated intensity of each particle identified, which was the sum of all of the pixel intensities within one particle as identified from the threshold image. The signal intensity values were stored in a matrix with each row representing a microbead and each column representing an image or time point in the experiment. The matrix was imported into MATLAB for further analysis.

As mentioned in section 4.3, control microbeads with fluorescent molecules directly attached to the microbeads were included in the imaging fields with the samples so that the baseline fluorescent intensity could be corrected for the signal loss from the photobleaching and temperature effects. In order to distinguish the control beads from the sample beads, the control image acquired with only sample beads present was processed by the same routine as the other images as described above. If one microbead identified in the first image was a control bead, then its signal intensity in the control image would be the sum of several pixels at the background intensity level because that bead was not present in the control image. This method was used to identify the control beads from the sample beads in the signal intensity matrix. Specifically, the background intensity level of the image was determined by calculating the mean intensity value in the image using the inverted threshold image, which eliminated all the bright spots. The maximum background bead intensity was

calculated by multiplying the mean background intensity level by the maximum pixel size allowed for a microbead. This background intensity value was used as a threshold to distinguish the sample beads from the control beads. Any microbead with signal larger than this threshold value was regarded as being present in the control image whereas the others were not. This method was used to distinguish control beads from sample beads by analyzing the control image. It was also used to identify microbeads that were washed away during the experiment by analyzing every image in the series. Those signals were eliminated from the signal intensity matrix because there was no signal in the later images

The signal intensity matrix was processed in MATLAB to plot the denaturation profiles and other statistics. After eliminating the signals from the microbeads that were washed away, the signals from the control beads were separated from those from the sample beads. The mean signal intensity in the control beads provided a top-line for the signal intensity level. The denaturation profiles were obtained by dividing the sample signal intensity by the mean signal intensity of the control beads at every time point. By plotting this signal over the temperature recorded at the time points when each image was taken, the denaturation profile was visualized. This denaturation signal was normalized by fitting the signal to a sigmoid curve and normalizing the top-line and baseline to one and zero, respectively. The melting temperature of the denaturation curve was determined as the temperature corresponding to a normalized signal intensity of 0.5.

4.4 Measurement of denaturation curves on the surface

Oligonucleotide templates were immobilized on a biotinylated cover slip for the measurement of denaturation curves on the surface. 75 × 50 mm^2 glass cover slips of 170 μm thickness were cleaned in 10% nitric acid for 30 minutes followed by washing with a large volume of deionized water and air dried after rinsing with acetone. The cover slips were then silanized with 2% 3-aminopropyl-triethoxysilane in 95% acetone (5% water) solution for 15 minutes so that the surface is derivatized with primary amino groups. Afterwards, the cover slips were washed three times in acetone, five minutes each, and then cured at 110ºC for 30 minutes. Next, the cover slips were derivatized with 5 mM NHS-PEG_{12}-Biotin in N,N-dimethylformamide containing 100 mM of triethylamine at room temperature for one hour and washed 4 times with acetone, 3 minutes each. Subsequently, the excess amino groups were blocked by acylation with 100 mM acetic anhydride in 1,4-dioxane with 100 mM triethylamine at room temperature for 30 minutes. After washing 4 times with acetone, 3 minutes each, and dried by blowing with filtered compressed air, the cover slips were stored under vacuum in a desiccator.

The biotinylated cover slip was assembled into the biochemical reaction chamber described in chapter 3.3. Double-sided tape (Scapa-702) with three 5 × 62 mm^2 channels was attached to a clean glass slide. Then, the biotinylated cover slip was attached to the double-sided tape to form a flow channel. The chamber assembly was baked in 110ºC overnight with kilograms of weight on top. The process removed the small bubbles trapped inside the type adhesive layer and ensured that the double-sided

tape was bonded to the both the slide and the cover slip to prevent leakage in the later steps. The subsequent steps were performed with the automated system.

The DNA samples were attached to microbeads which were immobilized on the surface for fluorescent imaging. The advantage of using microbeads is that tens to hundreds of thousands of DNA templates could be enriched into a small area in the image region resulting in higher fluorescent intensity. The same chemistry was used to attach the streptavidin-coated microbeads to the cover slips and to attach the biotinylated DNA templates onto the microbeads. The signal-to-noise ratio is increased dramatically because the surface chemistries only allowed the binding of microbeads but not DNA templates or probes, so the background signal due to non-specific binding of oligonucleotide probes is much lower intensity.

The protocol for preparing the DNA samples on the surface is the following. First, streptavidin coated 1 μm superparamagnetic beads (Dynal MyOne-C1) were suspended as a 0.0067% solution in 2X binding and washing buffer (B&W), which contains 2 M NaCl, 10 mM Tris-Cl, 2 mM EDTA, and 0.01% Triton X-100. The solution was injected into the flow chamber. A permanent magnet was dragged quickly underneath the flow channel to pull the magnetic microbeads down to the biotinylated cover slip surface. After incubation at 37ºC for 30 minutes to allow the streptavidin microbeads to bind effectively to the biotins on the cover slip surface, the magnet was dragged underneath the flow channel once again before the chamber was washed with 2 ml of wash buffer containing 33 mM phosphate buffer and 0.02% Triton X-100. Then, 400 nM of 5'-dual biotin-labeled oligonucleotide templates with

sequence "TACAGACTTAGTGGGGTAAAACTAGCATGACTGACTTCGTACA TGACTGATGGTCGATAC" were attached to the microbeads in 2X B&W buffer at 37ºC for 1 hour followed by washing with 2 ml of wash buffer. Next, 200 nM of oligonucleotide probes were hybridized to the templates attached to the microbeads in 2X B&W buffer at 37ºC for 1 hour. In the six experiments, each probe has the sequence of the first 21 to 27 bases of "CCATCAGTCATGTACGAAGTCAGTCAT" and an Alexa 546, 647, or 488 dye molecule attached to the 5' end respectively. The excess probes were washed away with at least 2 ml of wash buffer

When performing denaturation experiments through monitoring fluorescence, photobleaching and temperature effects on the fluorescent molecules may cause the fluorescent signal to decrease. In order to capture the true fluorescence decrease resulting from denaturation, we included controls in our experiment to correct for those effects. The same fluorescent molecules as used in the sample probes were used the controls. However, in the control, the molecules were covalently attached to the microbeads. The sequence of the control probes are the first 21 to 24 bases of "CCAT*CAGTCATGTACGAAGTCAGT", with biotin labeled on the 5' end and on the T marked with an asterisk, and an Alexa molecule labeled on the 3' end. 0.0125% microbeads and 62.5 nM of control oligonucleotide probes were mixed in 2X B&W buffer and incubated at 37ºC and agitated at 600 rpm for 1 hour. With this concentration of control probes, 20% of the biotin binding sites on the microbeads are occupied. This allows the microbeads to be further attached to the biotin surface in the reaction chamber in the later steps. The excess probes and the loosely-bound

probes were removed from the beads by washing the beads twice with 1 ml of 2X B&W buffer pre-heated to 65ºC. The control bead solution was resuspended in 1500 µl of 2X B&W buffer and vortexed before being dispensed into the flow chamber.

In each experiment, an image was taken to record where the sample beads were located before the control beads were flown into the chamber for immobilization. After the control beads were dispensed into the chamber, a permanent magnet was dragged underneath the channel to capture the control beads to the biotin surface. The control beads were incubated at 37ºC for 30 minutes before the loosely bound ones were washed away with 2 ml of wash buffer.

The denaturation curves of the hybridized oligonucleotide probes were measured by the change of fluorescence while the temperature was gradually increased in the chamber. The fluorescent intensity of the probes on the immobilized microbeads were monitored by acquiring 60 to 200 images at the same field of view over a period of time while the temperature increases linearly from 37ºC to 70ºC. A continuous wash to remove the denatured probes were maintained by an air-pressure driven flow of a wash buffer at a flow rate of 0.5~5 ml/min. The images were acquired using the same parameters over the course of one experiment. These parameters were chosen to utilize the full dynamic range of the CCD camera without saturation. The images were acquired using an EM gain of 4~20, and an exposure time of 40~200 ms, 30%, 50%, or 100% of power from the DG-5 light source, and a 20X/0.8NA plan-apochromat objective. The probes with Alexa 488, 546, or 647 were acquired using a multiband pass filter set (Semrock FF01-440/521/607/700-25), while the probes with

Alexa 750 were acquired using a Cy7 filter set (Chroma 41009). Before the images were acquired at each time point, fine autofocus was performed with 3~5 μm range and a step size of 0.25 μm at a neighboring spot. This ensures that the images were in focus while preventing photobleaching the samples due to the autofocus process. When the focus position was found to be at the edge of the range for two instances in a row, the focus position very likely drifted outside of the fine autofocus range, so a coarse autofocus over 20 μm was performed to maintain the correct focus position.

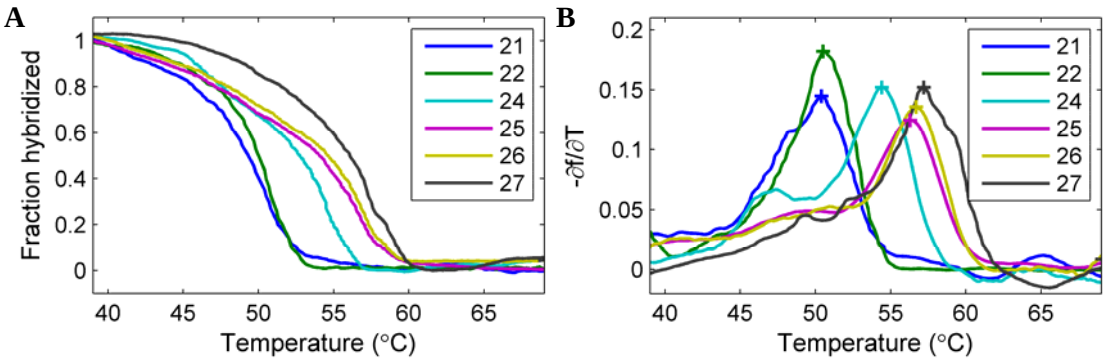


Figure 15. Denaturation profiles measured on the surface. (A) The denaturation profiles and **(B)** the negative derivative curves of 6 individual oligonucleotide probes measured on the surface. The sequences of the oligonucleotide probes contain the first 20 bases and additional 1 to 7 bases from the sequence "CCATCAGTCA TGTACGAAG TCAGTCAT" as indicated in the figure legend. The melting temperatures can be distinguished easily by the crosses marked at the peak of the negative derivative curves. The result demonstrates that denaturation profiles can be measured on the surface with single-base resolution.

The series of images acquired in each experiment was processed with the image processing and data analysis method described in section 4.3. The denaturation curves of six oligonucleotide probes are plotted in Figure 15A. The sequences of the oligonucleotide probes contain the first 20 bases and additional 1 to 7 bases from the

sequence “CCATCAGTCA TGTACGAAG TCAGTCAT” terminated at the 21, 22, 24, 25, 26, and 27 bases of this sequence, “CCATCAGTCATGTACGAAGTCA GTCAT”. In each experiment, the denaturation curves of one to three oligonucleotide probes with different fluorescent molecules labeled were measured. For each of the 21 to 27 base long oligonucleotide probes shown in Figure 15A, it is clear that the melting temperature of the shorter probe is lower than the one with one additional base. The data were smoothed and the negative derivatives were calculated numerically and smoothed. The results are shown in Figure 15B, with the peak of each negative derivative curve labeled with a cross for the easy visualization of the melting temperature. These results demonstrate that the denaturation profiles of oligonucleotide probes with lengths 21 to 27 can be measured on a surface with high accuracy. Single-base resolution was achieved by measuring the melting temperatures of these DNA fragments on the surface with our integrated system. Therefore, by measuring the denaturation profiles of labeled DNA fragments on a surface, we can read out the DNA sequence.

Some of the materials presented in this chapter are rearrangements from the publication: “DNA sequencing by denaturation: Principle and thermodynamic simulations”, Chen, Ying-Ja and Huang, Xiaohua, *Analytical Biochemistry*, 384(1): 170-179 (2009). The dissertation author was the primary investigator and author of this paper.

Chapter 5 Discussion and Future Work

5.1 Discussion and Future work

In this dissertation, the proof of principle for SBD has been demonstrated in theory and by experimentally measuring denaturation curves of oligonucleotide probes on the surface. However, when using SBD for sequencing, there may be complications in the Sanger reaction due to the surface environment. Further studies are required to determine the optimal parameters or modifications in order to perform Sanger reaction efficiently and uniformly on the surface.

When incorporating the sequencing chemistry from SBD with the upstream sample preparation and amplification methods, SBD offers great flexibility and compatibility due to its simple chemistry. Densely-packed arrays of beads with nearly 100% filling could be used to deposit the sequencing templates on the surface to achieve a much higher imaging efficiency and throughput [50]. This will be a great improvement over the randomly-distributed microbeads used in the current study.

Another method for preparing templates is to use rolling circle amplification (RCA) to amplify from circular templates. The long linear products from RCA will assume a compact sphere-like structure with a characteristic Flory radius, which when functionalized using modified RCA primers can be deposited onto the array to replace the microbeads. These RCA DNA molecules are not only more compact, but also provide solution-like environments that are more favorable for DNA polymerases compared to the surface environment on a bead. Furthermore, the fluorescent signal

will be more concentrated because the whole volume of the RCA-amplified template will contribute to the signal rather than only the surface of a bead. When performing SBD on these templates that are tandem repeats in nature, the Sanger reaction length can be limited because at the end of each strand the primer from the next strand blocks further synthesis. This can improve the accuracy of SBD because fragments beyond the read length of SBD are not produced.

In SBD, the signal is detected from the denaturation event of a fluorescently-labeled Sanger reaction product. The sequence is determined by discriminating the denaturation of one Sanger fragment from the fragment with one additional base. Although the fragment with one additional base is usually more stable due to its two or three additional hydrogen bonds and base-stacking, its stability could be complicated by the interaction between the labeled fluorescent molecule and the following bases on the template. It has been shown that fluorescent molecules may interact with the bases on the DNA resulting in an extra stabilizing factor [51], which may alter the denaturation order of one fragment from another depending on the base sequences. Further studies are needed to fully characterize these interactions. The results would be useful in determining correction methods after a read is sequenced. When later sequences are being read, one could predict if an inversion of denaturation event caused by such dye-base interaction was present and re-call the sequence by taking this into account.

The potential sources of sequencing errors have been investigated. The majority of the errors is generated during the base-calling process and result from the

imperfect fitting of the negative derivatives of the denaturation curves to a sum of Gaussian curves. DNA sequences with complex secondary structures, palindromes, or tandem repeats may cause unexpected denaturation properties, which could lead to errors. The issues are more pronounced in certain sequences where a short repeat of one base type with lower binding strengths (A & T) is intercepted by another short repeat. For example, a sequence with "ATTAA" composition is more likely to produce an error with our calling algorithm. In addition, the denaturation event is subject to cooperativity and the negative derivative of a denaturation profile may be slightly skewed and not represented precisely with a Gaussian curve. Another major source of base-calling errors results from the difficulty in resolving the more extensive overlaps between the denaturation curves of longer fragments. An increase in read accuracy can be achieved by limiting the Sanger reaction to a certain length either by using a high concentration of dideoxyribonucleotides or terminating the reaction with a primer pre-hybridized a short distance from the sequencing primer. When terminating the reaction at a defined length, the interference from longer Sanger fragments is eliminated. This results in a clearer determination of the last few bases to sequence.

In the fitting process, we assume that each Sanger fragment is uniformly populated and its denaturation profile follows a normal distribution. In real experiments each Sanger fragment may be differentially represented during the sequencing reaction. However, this problem can be alleviated to some degree by using engineered DNA polymerases such as Sequenase version 2.0 or Thermo Sequenase to

generate more uniformly represented DNA fragments. These enzymes have been shown to produce uniform bands in Sanger sequencing since they do not discriminate between dideoxyribonucleotides and deoxyribonucleotides and have much less sequence dependency [52-54]. The base-calling algorithm appears to be robust against potential noise in the measurement. We want to emphasize that the simple algorithm described here is sufficient to illustrate the principle of SBD but obviously is not the best one. Higher sequencing accuracy can be achieved by further improvement in the base-calling algorithm, for examples, by optimizing the parameters with experimental data or by replacing the Gaussian fit with numerical methods such as higher-order derivatives to determine the component Sanger fragments. The effects of fluorescent labels and dangling ends of the oligonucleotides on the thermodynamic properties of the DNA in the denaturation process are more complex and have not been considered in our simulations. Further work is required to develop a more robust and better base-calling algorithm to reduce the base-calling errors. Unlike DNA hybridization processes where kinetic factors play important roles and are difficult to control, the denaturation process is strictly determined by the thermodynamic properties of the double-stranded DNA molecules and is more predictable. We believe that highly accurate sequencing can be feasible with SBD.

When the template to sequence has a low GC content, more errors can occur for the following reason. As shown by the simulations using the nearest-neighbor parameters to estimate the melting profile of DNA strands, the change in melting temperature is smaller when there are more A-T pairs due to its lower bond energy

from only two hydrogen bonds rather than three. Experimentally, this issue can be resolved by adding a chemical such as tetramethylammonium chloride which is known to interact differentially with the A-T base pairs and increases their melting temperature to be the same as that of G-C base pairs. In the presence of such reagents, the effect of the base composition on melting temperature is neutralized so that the T_m is dependent only on the length of the DNA [55]. This will even out the melting curves and potentially eliminate the majority of the errors in SBD.

In this study, the read length of SBD is estimated to be around 20 bases. This short read length is due to the limited resolution between the melting temperatures of longer DNA fragments. Besides increasing the sensitivity of the imaging system to increase the signal-to-noise ratio, the number of fragments being measured can also have a significant effect on the accuracy and resolution of SBD. It is worth studying the potential resolution of the measurement of melting temperature as the number of templates used. In addition, the read length of SBD can be extended to 40-50 bases experimentally by using a primer which can be cleaved off at the 3' end of the first sequencing primer by, for example, using photochemically cleavable linkers [56, 57]. Even though the maximum read length is limited in SBD, with a potential read length of 40-50 bases, it can be used for genome re-sequencing and other applications where a short read is sufficient to identify the sequence unambiguously.

5.1.1 Imaging speed and throughput of SBD

For DNA sequencing technologies that use fluorescence detection, the limitation of the speed for sequencing is the speed for imaging. Therefore, it is

important that we build an imaging system that provides the maximum speed and throughput for SBD. Here we calculate the imaging speed and throughput of SBD using the system constructed as described in Chapter 3. The equations used for the calculation are listed in Table 1.

Table 1. List of equations for calculating the throughput for SBD.

Area imaged in one field: $$1\text{Mpixel}\times\left(\frac{8\mu\text{m/pixel}}{10\text{X}}\right)^2 = 6.4\times10^5\,\mu\text{m}^2$$
Number of fields per flow cell: $$\frac{2\text{cm}\times 6\text{cm}}{6.4\times10^5\,\mu\text{m}^2} = 1875\text{ fields}$$
Time to image one field: $$100ms + 4\times 40ms = 260ms\,/\,field$$
Number of sample points to acquire for SBD detection: $$\frac{(70-20)°C}{0.5°C} = 100$$
Time to image the entire chamber: $$1875\,fields\times 260ms\,/\,field\times 100samples = 4.9\times10^4\,s = 813\,\text{min} = 13.5hrs$$
Number of reads in one field: $$\frac{1\times10^6\,pixels}{9\,pixels\,/\,feature} = 1.1\times10^5\,reads\,/\,field$$
Number of bases sequenced per run: $$20\ bases/read\times 1.1\times10^5\ reads/fields\times 1875\ fields = 4.2\times10^9\ bases$$

In SBD, many images are required to measure the ensemble denaturation profiles of the Sanger fragments as the temperature is gradually increased. In

order to account for each denaturation events, the interval between each sample point taken must be at least smaller than the difference in melting temperature between fragments with single base differences. From our simulations study, the average difference in melting temperature between fragments with single base difference is about 2.5ºC for oligonucleotides of 12 to 32 bases. However, the difference in melting temperature decreases dramatically when more bases are added and drops to around 1ºC when the fragment length is 30 bases. Therefore, we reason that it is practical to take a measurement each time the temperature is increased by 0.3~1ºC. Therefore, 50~150 images must be acquired for each channel if the measurements are taken over a temperature range from 20 to 70ºC,

Now we calculate the amount of time required to acquire each measurement. For each sample, images for 4 colors are acquired. By controlling the excitation intensities, the minimum exposure time of 38 ms can be achieved. This is limited by the data readout rate of the camera which has a maximum frame rate of 35 MHz without binning. With a DG-5 light switching device, only 1 ms is needed to switch between the channels so that the camera can continuously acquire images. As a result, it takes roughly 160 ms to acquire the 4-color images for one measurement. For high-throughput applications, it may require to scan through many imaging fields. If a 20X or a 10X objective is used, the time required for the linear encoder stage to move to a neighboring field is about 100 ms. Therefore, the total time required for imaging each field is about 260 ms.

In SBD, the samples are deposited onto an array to be imaged on a surface. Although in the experiments presented here the samples were randomly distributed, the samples can be deposited on an ordered array to fully utilize the surface area and maximizing imaging efficiency [50]. By using an array of samples with 800 nm in size and 1.2 μm in pitch and using a 20X objective, and by aligning the array to the pixels on the EMCCD camera that are 8 μm in size, it is feasible to image one sample with an array of 3×3 pixels where the signal from each sample is captured within 2×2 pixels while the surrounding 5 pixels are dark helping to distinguish the neighboring samples. With the 1004×1002 pixel EMCCD camera used, the signals from a total of 111,556 samples can be detected in one field of view. If a 10X objective and arrays of 800 nm in size and 1.6 μm in pitch are used, only 2x2 pixels are required to image each sample and 0.25 million of samples can be imaged in one EMCCD imaging field. If further miniaturization of the array is desired to align each sample to one pixel on the camera, then every pixel will be bright. In this case, there must be patterns built into the array for the alignment of pixel to samples. These patterns can be ordered patterns that are present in the array as dark spots in every field of view. With the minimal dark patterns embedded in the array, nearly one million samples can be imaged in one EMCCD imaging field. These arrays can be fabricated through photolithography or nanolithography technologies.

In addition to the large number of reads that can be generated from the one hundred thousand samples in one field of view, the sequencing throughput

will be higher by scanning through many fields. The imaging surface of the system we constructed is 20×60 mm^2. The imaging area on the camera is the pixel size of 8 μm multiplied by the number of pixels (1004×1002) in the camera, which is 8064×8016 μm^2. By using a 20X or a 10X objective, the observed area on the specimen in one field of view becomes 403.2×400.8μm^2 or 806.4×801.6 μm^2. Therefore, the total number of fields in the surface is 1875. By scanning through this area, 209 million reads can be imaged. Since the time required to image one field in each channel and to move from one field to the next is 260 ms, it takes 487.5 s or 8 minutes to image the entire flow chamber.

As mentioned above, SBD requires the acquisition of 50~150 images per run for the complete measurement of a denaturation curve. When imaging in only one single field, 160 ms are required for imaging at each temperature point, which results in a total of 8~24 s for the entire denaturation process. When scanning through the entire flow chamber in 4 color channels, this process will take 7~20 hours.

The principle of SBD depends on the separation in melting temperatures between DNA fragments with single-base differences. Since the difference in melting temperature of longer DNA fragments become indistinguishable, we estimate that the raw read length of SBD is 20 bases for each template. With the 209 million templates on the array, 2.2 million bases can be read in one field, and 4.2 billion bases can be read from the 20 × 60 mm^2 flow chamber. Therefore, the average throughput of SBD is estimated to be 210~600 million bases per hour.

This will give the the equivalent of 1~3X coverage of a human genome in less than a day.

5.1.2 Estimated cost of sequencing with SBD

As described above, SBD employs simple chemistry and only requires the denaturation of the Sanger fragments. This eliminates the need for large quantity of reagents. Here we calculate an estimated cost to show that SBD can be used to bring the cost of genome sequencing down significantly.

The volume of the flow cell is 144 µl. For the estimations below, we assume that there will be around 10^5 copies of template on each feature and 200 million features, which will be 2×10^{13} molecules or 1.7×10^{-11} moles. The Sanger sequencing reaction in the flow cell will require DNA polymerases, deoxynucleotides, fluorescently-labeled dideoxynucleotides, primers, and buffer. The amount of polymerase required would be about 10 units for the 144 µl of reaction volume. The cost of Sequenase or Thermo Sequenase is about $220 for 325 units (USB Corp.), which will be about $7 for SBD. The nucleotide needed for SBD will be around 140 nmoles for a 100X excess in reaction on the 200 million fragments in the flow cell with an estimated read length of 40. Since a tube of 4 × 25 µmol is around $200, the cost of nucleotides for SBD will be around $1. The fluorescently-labeled dideoxynucleotides are the most expensive reagents that will be used. They can be purchased as sequencing reaction kits costing $270 for 24 reactions (Applied Biosystems BigDye Terminator v1.1

Cycle Sequencing Kit). It is estimated that the equivalent of 3 reaction reagents are required for SBD, which is about $35. The primers used for sequencing are often supplied in the kit, but if synthesized separately will be about $5 with a 25 nmole custom oligonucleotide synthesis of 12~20 bases long primer. In some cases, single-stranded DNA binding protein (SSB) is required, with 10^5 copies of 200 million templates and a 10-fold excess of SSB used, roughly 20 µl from a 2 mg/ml vial stock is required (EPICENTRE Biotechnologies, $104 for 200 µg), which will cost about $10. The buffer for the Sanger reaction are often supplied with the polymerase or nucleotide kits or their ingredients such as salt compounds can be purchased in bulk, which averages much less than $1 for the whole reaction in SBD. By adding the cost from all of the above reagents for performing a sequencing reaction with SBD, the reagent cost is about $60.

In addition to reagent cost, sample preparation and amplification contributes greatly to the total cost for sequencing. SBD can be compatible with a number of different sample preparation methods. In general, the attachment of adaptor oligonucleotides will be necessary for SBD, which as estimated above costs $10. Here we estimate the cost of amplification using emulsion PCR. Since SBD templates are attached to 1-µm microbeads, micro-emulsions of 10 µm in diameter can be used for the emulsion PCR reaction. The volume in such emulsion is 1.7×10^{-13} liters. With 200 million templates and 10% efficiency, the total volume of reagents necessary for emulsion PCR of SBD templates is 340 µl. As estimated above, the cost of DNA polymerase for 340 µl of reaction is around

$12. Similarly, the nucleotides cost about $2, and the primers about $10. The mineral oil and surfactant used in emulsion PCR can be purchased in bulk, which averages much less than $1 for the entire reaction. All of these reagents will cost about $35.

The above calculation accounted for most of the components that contribute to the cost for SBD from sample preparation to sequencing, which amount to around $100. However, for the accuracy necessary to re-sequence a genome, we estimate that a 10-fold coverage is required. This increases the cost by 10 fold to about $1000. The cost estimation provides an idea of the order of magnitude in cost for genome sequencing with SBD.

5.2 Conclusion

In this dissertation, I present the theoretical basis and experimental proof of principle of a new DNA sequencing technology called sequencing by denaturation (SBD). The theoretical basis of SBD was described in Chapter 2. The denaturation profiles of the DNA fragments generated by fluorescently-labeled dideoxyribonucleotides were simulated by melting curve analysis using thermodynamic principles and data. The simulations showed that melting temperatures of oligonucleotides increase monotonously as each additional base is added. Using this property, an algorithm for base-calling was developed to decode the DNA sequence from the intensity data. The average error rates versus read lengths were estimated. A high-speed fluorescent imaging system integrating a fluorescent microscope with

temperature control and fluidics through a custom-built reaction chamber was presented in Chapter 3. This system was fully characterized so that it could be customized for many applications where a high-throughput fluorescence detection system is required. This system was programmed for the experimentation for SBD and the results are presented in Chapter 4. In Chapter 4, denaturation profiles of many fluorescently-labeled oligonucleotide probes were measured first in solution to demonstrate that the base-calling algorithm developed previously can correctly determine the base sequence of a DNA template. These measurements were later performed on the surface in a high-throughput manner. An image processing routine was established to analyze a series of images acquired over time during an SBD experiment. The results showed that denaturation profiles could be measured on the surface and that the difference in melting temperature between oligonucleotides with single-base differences could be distinguished. We calculated the throughput of this system to show its potential for genome sequencing. In summary, we have demonstrated with theoretical simulations and experiments that SBD is a feasible method for high-throughput DNA sequencing and we have also constructed the instrumentation to perform SBD.

In Table 2 we compare SBD to the next-generation sequencing platforms that are currently available. Despite the different sequencing chemistries used, the read lengths of most of these technologies are short, but each of them achieves a similar throughput of around 0.5 to 1 billion bases per day. As estimated in section 5.1.1, the throughput of SBD can potentially be around 10 billion bases per day, which is about

10 times higher. This improvement is largely due to the scalability of the method and the array technologies that are emerging. We understand that the SBS and SBL technologies are also very scalable and may be able to reach such improvement in throughput as well.

Table 2. Comparison of SBD to other available technologies.

	454 /Roche GS FLX	ABI SOLiD	Illumina/Solexa 1 G Analyzer	Our method
Chemistry	SBS (pyroseq.)	SBL	SBS	SBD
Read length (bp)	400	35 25 x 2	36	20
Throughput (bp/day)	1 B	600 M	600 M	5~15 B
Cost per genome	$1M	N/A ($100K Church Lab)	N/A (>$100K)	~$1K
Run time	500 Mbp / 10 hrs	3~4 Gbp / 6 days 5~6 Gbp / 10 days	1.5 Gbp / 2.5 days	4.2 Gbp / 7~20 hrs

The cost of re-sequencing a human genome for SBD as estimated in section 5.1.2 is less than 1000 US dollars, 2-3 orders of magnitude lower than the cost for other technologies. This mainly results from the reduction in reagents. Since SBD chemistry only requires denaturation through heating and washing, no expensive reagents need to be delivered to the flow cell in every cycle as compared to all other available technologies. This reduces the cost of reagents by roughly 100 fold. The cost

for genome sequencing is also contributed by other factors such as sample preparation and amplification. Because SBD is a highly scalable method, the volume of reagents required for both sample preparation and amplification can be maintained at very low levels. We estimated that the cost of re-sequencing a genome by SBD can be significantly lower than other technologies. With these advantages, SBD is a very appealing technology for DNA sequencing.

In SBD, multiple sequencing runs could be performed on the same templates in a flow cell, perhaps with a set of primers of different lengths, to significantly improve sequencing accuracy. A single run in our system can potentially produce 4.2 Gbp of data from 200 million reads within a day. These capabilities are highly compatible with applications such as genome resequencing, high-throughput SNP genotyping and digital analysis of gene expression when only a short read length is required to uniquely determine the identity of the fragment and large numbers of reads are desired. With a competitive throughput and orders of magnitude reduction in cost, we believe that SBD can be a promising sequencing technology for a wide variety of applications.

Some of the materials presented in this chapter are rearrangements from the publication: “DNA sequencing by denaturation: Principle and thermodynamic simulations”, Chen, Ying-Ja and Huang, Xiaohua, *Analytical Biochemistry*, 384(1): 170-179 (2009). The dissertation author was the primary investigator and author of this paper.

References

1. J. Shendure, R.D. Mitra, C. Varma, and G.M. Church, Advanced sequencing technologies: methods and goals. *Nat Rev Genet* 5: 335-344 (2004).
2. M.L. Metzker, Emerging technologies in DNA sequencing. *Genome Res* 15: 1767-1776 (2005).
3. D. Hanahan, and R.A. Weinberg, The hallmarks of cancer. *Cell* 100: 57-70 (2000).
4. F. Sanger, S. Nicklen, and A.R. Coulson, DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 74: 5463-5467 (1977).
5. L.M. Smith, S. Fung, M.W. Hunkapiller, T.J. Hunkapiller, and L.E. Hood, The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic Acids Res* 13: 2399-2412 (1985).
6. X.C. Huang, M.A. Quesada, and R.A. Mathies, DNA sequencing using capillary array electrophoresis. *Anal Chem* 64: 2149-2154 (1992).
7. D. Dressman, H. Yan, G. Traverso, K.W. Kinzler, and B. Vogelstein, Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc Natl Acad Sci U S A* 100: 8817-8822 (2003).
8. R.D. Mitra, and G.M. Church, In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Res* 27: e34 (1999).
9. R.D. Mitra, J. Shendure, J. Olejnik, O. Edyta Krzymanska, and G.M. Church, Fluorescent in situ sequencing on polymerase colonies. *Anal Biochem* 320: 55-65 (2003).
10. D.R. Bentley, S. Balasubramanian, H.P. Swerdlow, G.P. Smith, J. Milton, C.G. Brown, et al., Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456: 53-59. (2008).
11. A. Marziali, and M. Akeson, New DNA sequencing methods. *Annu Rev Biomed Eng* 3: 195-223 (2001).
12. C.A. Emrich, H. Tian, I.L. Medintz, and R.A. Mathies, Microfabricated 384-lane capillary array electrophoresis bioanalyzer for ultrahigh-throughput genetic analysis. *Anal Chem* 74: 5076-5083 (2002).
13. T.S. Seo, X. Bai, H. Ruparel, Z. Li, N.J. Turro, and J. Ju, Photocleavable fluorescent nucleotides for DNA sequencing on a chip constructed by site-specific coupling chemistry. *Proc Natl Acad Sci U S A* 101: 5488-5493 (2004).
14. Z. Li, X. Bai, H. Ruparel, S. Kim, N.J. Turro, and J. Ju, A photocleavable fluorescent nucleotide for DNA sequencing and analysis. *Proc Natl Acad Sci U S A* 100: 414-419 (2003).

15. J. Guo, N. Xu, Z. Li, S. Zhang, J. Wu, D.H. Kim, et al., Four-color DNA sequencing with 3' O -modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proc Natl Acad Sci USA* 105: 9145-9150 (2008).

16. M. Margulies, M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, et al., Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376-380 (2005).

17. M. Ronaghi, Pyrosequencing sheds light on DNA sequencing. *Genome Res* 11: 3-11 (2001).

18. M. Chee, R. Yang, E. Hubbell, A. Berno, X.C. Huang, D. Stern, et al., Accessing genetic information with high-density DNA arrays. *Science* 274: 610-614 (1996).

19. N. Patil, A.J. Berno, D.A. Hinds, W.A. Barrett, J.M. Doshi, C.R. Hacker, et al., Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294: 1719-1723 (2001).

20. A. Pihlak, G. Bauren, E. Hersoug, P. Lonnerberg, A. Metsis, and S. Linnarsson, Rapid genome sequencing with short universal tiling probes. *Nat Biotechnol* 26: 676-684 (2008).

21. J. Shendure, G.J. Porreca, N.B. Reppas, X. Lin, J.P. McCutcheon, A.M. Rosenbaum, et al., Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309: 1728-1732 (2005).

22. J.B. Kim, G.J. Porreca, L. Song, S.C. Greenway, J.M. Gorham, G.M. Church, et al., Polony Multiplex Analysis of Gene Expression (PMAGE) in Mouse Hypertrophic Cardiomyopathy. *Science* 316: 1481-1484 (2007).

23. N. Cloonan, A.R. Forrest, G. Kolle, B.B. Gardiner, G.J. Faulkner, M.K. Brown, et al., Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* 5: 613-619 (2008).

24. I. Braslavsky, B. Hebert, E. Kartalov, and S.R. Quake, Sequence information can be obtained from single DNA molecules. *Proc Natl Acad Sci U S A* 100: 3960-3964 (2003).

25. J. Korlach, A. Bibillo, J. Wegener, P. Peluso, T.T. Pham, I. Park, et al., Long, processive enzymatic DNA synthesis using 100% dye-labeled terminal phosphate-linked nucleotides. *Nucleosides Nucleotides Nucleic Acids* 27: 1072-1083. (2008).

26. J. Korlach, P.J. Marks, R.L. Cicero, J.J. Gray, D.L. Murphy, D.B. Roitman, et al., Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proc Natl Acad Sci USA* 105: 1176-1181 (2008).

27. M.J. Levene, J. Korlach, S.W. Turner, M. Foquet, H.G. Craighead, and W.W. Webb, Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* 299: 682-686 (2003).

28. T.D. Harris, P.R. Buzby, H. Babcock, E. Beer, J. Bowers, I. Braslavsky, et al., Single-molecule DNA sequencing of a viral genome. *Science* 320: 106-109. (2008).

29. I. Rasnik, S.A. McKinney, and T. Ha, Nonblinking and long-lasting single-molecule fluorescence imaging. *Nat Methods* 3: 891-893 (2006).

30. H. Wang, and D. Branton, Nanopores with a spark for single-molecule detection. *Nat Biotechnol* 19: 622-623 (2001).

31. D. Branton, D.W. Deamer, A. Marziali, H. Bayley, S.A. Benner, T. Butler, et al., The potential and challenges of nanopore sequencing. *Nat Biotechnol* 26: 1146-1153. (2008).

32. R. Drmanac, I. Labat, I. Brukner, and R. Crkvenjakov, Sequencing of megabase plus DNA by hybridization: theory of the method. *Genomics* 4: 114-128 (1989).

33. Z. Strezoska, T. Paunesku, D. Radosavljevic, I. Labat, R. Drmanac, and R. Crkvenjakov, DNA sequencing by hybridization: 100 bases read by a non-gel-based method. *Proc Natl Acad Sci USA* 88: 10089-10093 (1991).

34. R. Drmanac, S. Drmanac, Z. Strezoska, T. Paunesku, I. Labat, M. Zeremski, et al., DNA-Sequence Determination by Hybridization - a Strategy for Efficient Large-Scale Sequencing. *Science* 260: 1649-1653 (1993).

35. S. Drmanac, D. Kita, I. Labat, B. Hauser, C. Schmidt, J.D. Burczak, et al., Accurate sequencing by hybridization for DNA diagnostics and individual genomics. *Nat Biotechnol* 16: 54-58 (1998).

36. R. Drmanac, S. Drmanac, G. Chui, R. Diaz, A. Hou, H. Jin, et al., Sequencing by hybridization (SBH): advantages, achievements, and opportunities. *Adv Biochem Eng Biotechnol* 77: 75-101 (2002).

37. T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286: 531-537 (1999).

38. N. von Ahsen, Oellerich, M., Armstrong, V.W. and Schutz, E., Application of a thermodynamic nearest-neighbor model to estimate nucleic acid stability and optimize probe design: prediction of melting points of multiple mutations of apolipoprotein B-3500 and factor V with a hybridization probe genotyping assay on the LightCycler. *Clin Chem* 45: 2094-2101 (1999).

39. C.D. Bennett, M.N. Campbell, C.J. Cook, D.J. Eyre, L.M. Nay, D.R. Nielsen, et al., The LightTyper: high-throughput genotyping using fluorescent melting curve analysis. *Biotechniques* 34: 1288-1292, 1294-1285 (2003).

40. E. Lyon, Mutation detection using fluorescent hybridization probes and melting curve analysis. *Expert Rev Mol Diagn* 1: 92-101 (2001).

41. P.N. Borer, Dengler, B., Tinoco, I., Jr. and Uhlenbeck, O.C., Stability of ribonucleic acid double-stranded helices. *Journal of Molecular Biology* 86: 843-853 (1974).

42. K.J. Breslauer, R. Frank, H. Blocker, and L.A. Marky, Predicting DNA duplex stability from the base sequence. *Proc Natl Acad Sci USA* 83: 3746-3750 (1986).

43. R.A. Dimitrov, and M. Zuker, Prediction of hybridization and melting for double-stranded nucleic acids. *Biophys J* 87: 215-226 (2004).

44. J. SantaLucia, Jr., A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci USA* 95: 1460-1465 (1998).

45. J. SantaLucia, Jr., H.T. Allawi, and P.A. Seneviratne, Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry* 35: 3555-3562 (1996).

46. N. Sugimoto, S. Nakano, M. Yoneyama, and K. Honda, Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res* 24: 4501-4505 (1996).

47. B.H. Zimm, and J.K. Bragg, Theory of the Phase Transition between Helix and Random Coil in Polypeptide Chains. *Journal of Chemical Physics* 31: 526-531 (1959).

48. R. Owczarzy, I. Dunietz, M.A. Behlke, I.M. Klotz, and J.A. Walder, Thermodynamic treatment of oligonucleotide duplex-simplex equilibria. *Proc Natl Acad Sci USA* 100: 14840-14845 (2003).

49. A. Panjkovich, and F. Melo, Comparison of different melting temperature calculation methods for short DNA sequences. *Bioinformatics* 21: 711-722 (2005).

50. K.D. Barbee, and X. Huang, Magnetic assembly of high-density DNA arrays for genomic analyses. *Anal Chem* 80: 2149-2154. (2008).

51. B.G. Moreira, Y. You, M.A. Behlke, and R. Owczarzy, Effects of fluorescent dyes, quenchers, and dangling ends on DNA duplex stability. *Biochem Biophys Res Commun* 327: 473-484 (2005).

52. M.A. Reeve, and C.W. Fuller, A novel thermostable polymerase for DNA sequencing. *Nature* 376: 796-797 (1995).

53. C.W. Fuller, B.F. McArdle, A.M. Griffin, and H.G. Griffin, DNA sequencing using sequenase version 2.0 T7 DNA polymerase. *Methods Mol Biol* 58: 373-387 (1996).

54. S. Kumar, C.W. Fuller, S. Nampalli, M. Khot, I. Livshin, L. Sun, et al., Uniform band intensities in fluorescent dye terminator sequencing. *Nucleosides Nucleotides* 18: 1101-1103 (1999).

55. M.L.M. Anderson, Nucleic Acid Hybridization, *BIOS Scientific Publishers Limited*, Oxford, UK, 1999.

56. J. Olejnik, H.C. Ludemann, E. Krzymanska-Olejnik, S. Berkenkamp, F. Hillenkamp, and K.J. Rothschild, Photocleavable peptide-DNA conjugates: synthesis and applications to DNA analysis using MALDI-MS. *Nucleic Acids Res* 27: 4626-4631 (1999).

57. P.M. Vallone, K. Fahr, and M. Kostrzewa, Genotyping SNPs using a UV-photocleavable oligonucleotide in MALDI-TOF MS. *Methods Mol Biol* 297: 169-178 (2005).