Method Development and Applications of Pyrosequencing Technology

Baback Gharizadeh

Royal Institute of Technology
Department of Biotechnology

Stockholm 2003

ISBN 91-7283-609-1

Abstract

The ability to determine nucleic acid sequences is one of the most important platforms for the detailed study of biological systems. Pyrosequencing technology is a relatively novel DNA sequencing technique with multifaceted unique characteristics, adjustable to different strategies, formats and instrumentation. The aims of this thesis were to improve the chemistry of the Pyrosequencing technique for increased read-length, enhance the general sequence quality and improve the sequencing performance for challenging templates. Improved chemistry would enable Pyrosequencing technique to be used for numerous applications with inherent advantages in accuracy, flexibility and parallel processing.

Pyrosequencing technology, at its advent, was restricted to sequencing short stretches of DNA. The major limiting factor was presence of an isomer of dATP[S], a substitute for the natural dATP, which inhibited enzyme activity in the Pyrosequencing chemistry. By removing this non-functional nucleotide, we were able to achieve DNA read-lengths of up to one hundred bases, which has been a substantial accomplishment for performance of different applications. Furthermore, the use of a new polymerase, called Sequenase, has enabled sequencing of homopolymeric T-regions, which are challenging for the traditional Klenow polymerase. Sequenase has markedly made possible sequencing of such templates with synchronized extension.

The improved read-length and chemistry has enabled additional applications, which were not possible previously. DNA sequencing is the gold standard method for microbial and vial typing. We have utilized Pyrosequencing technology for accurate typing of human papillomaviruses, and bacterial and fungal identification with promising results.

Furthermore, DNA sequencing technologies are not capable of typing of a sample harboring a multitude of species/types or unspecific amplification products. We have addressed the problem of multiple infections/variants present in a clinical sample by a new versatile method. The multiple sequencing primer method is suited for detection and typing of samples harboring different clinically important types/species (multiple infections) and unspecific amplifications, which eliminates the need for nested PCR, stringent PCR conditions and cloning. Furthermore, the method has proved to be useful for samples containing subdominant types/species, and samples with low PCR yield, which avoids re-performing unsuccessful PCRs. We also introduce the sequence pattern recognition when there is a plurality of genotypes in the sample, which facilitates typing of more than one target DNA in the sample. Moreover, target specific sequencing primers could be easily tailored and adapted according to the desired applications or clinical settings based on regional prevalence of microorganisms and viruses.

Pyrosequencing technology has also been used for clone-checking by using pre-programmed nucleotide addition order, EST sequencing and SNP analysis, yielding accurate and reliable results.

Keywords: apyrase, bacterial identification, dATP[S], EST sequencing, fungal identification, human papillomavirus (HPV), microbial and viral typing, multiple sequencing primer method, Pyrosequencing technology, Sequenase, single-stranded DNA-binding protein (SSB), SNP analysis

©Baback Gharizadeh
To my parents
and to my son, Jonathan
“The stars are beautiful, because of a flower that cannot be seen”

“What is essential is invisible to the eye”

*Le Petit Prince* (The Little Prince)
Antoine-Marie-Roger de Saint-Exupéry
This thesis is based on the following publications, which in the text will be referred to by their roman numerical:

List of publications


## Table of contents

1. Introduction ........................................................................................................... 1

2. Polymerase chain reaction ...................................................................................... 3

3. DNA sequencing techniques .................................................................................. 5
   3.1. Dideoxy DNA sequencing by chain termination .............................................. 5
       3.1.1. Slab gel electrophoresis ............................................................................ 6
       3.1.2. Capillary electrophoresis ......................................................................... 7
   3.2. Non-electrophoretic DNA sequencing methods .............................................. 7
       3.2.1. Mass spectrometry ................................................................................... 7
       3.2.2. Sequencing-by-hybridization ................................................................. 8
       3.2.3. Sequencing-by-synthesis ......................................................................... 8

4. Pyrosequencing technology .................................................................................... 12
   4.1. Pyrosequencing by the solid-phase approach ................................................. 12
   4.2. Pyrosequencing by the liquid-phase approach ................................................. 15
   4.3. Template preparation for the Pyrosequencing method .................................... 19
       4.3.1. Single-strand approach ........................................................................... 19
       4.3.2. Double-strand approach ........................................................................ 20

5. Present investigation ............................................................................................... 22
   5.1. Methodological improvements of Pyrosequencing technology ..................... 22
       5.1.1. Read-length improvement (Paper I) ......................................................... 22
       5.1.2. Advancements in sequencing T-homopolymeric regions (Paper II) .......... 24
       5.1.3. Improved DNA sequencing by SSB (Paper VIII) ..................................... 26
       5.1.4. Pre-programmed DNA sequencing (Paper III) ....................................... 28
       5.1.5. Multiple sequencing-primer method and applications (Papers IV and V) ... 29
           5.1.5.1. Detection of multiple infections (Paper IV) ....................................... 30
           5.1.5.2. Unspecific amplification products (Paper IV) .................................. 36
           5.1.5.3. Subdominant types and amplicons with low yield ......................... 38
           5.1.5.4. Faster and more efficient DNA-sequencing (Paper V) .................... 39
   5.2. Applications of Pyrosequencing technology (Papers V-IX) ............................... 42
       5.2.1. Microbial and viral typing (Papers V-VII) ............................................... 42
           5.2.1.1. Human papillomavirus genotyping (Paper VI) ............................... 42
           5.2.1.2. Bacterial identification (Paper V) ..................................................... 44
           5.2.1.3. Fungal identification (Paper VIII) ................................................... 47
       5.2.2. EST sequencing (Paper VIII) .................................................................. 49
       5.2.3. SNP analysis (Paper IX) ......................................................................... 50
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6. Future trends of Pyrosequencing technology</td>
<td>52</td>
</tr>
<tr>
<td>7. Concluding remarks</td>
<td>53</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>54</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>56</td>
</tr>
<tr>
<td>References</td>
<td>58</td>
</tr>
</tbody>
</table>

Original papers (I-IX)
1. Introduction

Recent impressive advances in DNA sequencing technologies have accelerated the detailed analysis of genomes from many organisms. We have been observing reports of complete or draft versions of the genome sequence of several well-studied, multicellular organisms. Human biology and medicine are in the midst of a revolution by Human Genome Project (Yager et al., 1991) as the main catalyst.

Chromosomal DNA contains the blueprint of cellular structure and function in specific regions called genes. The double-stranded structure of DNA was revealed by Watson and Crick in 1953 (Watson and Crick, 1953) and consequently, the mechanism of translation from DNA to proteins was found a few years later. The regulation and content of genetic information are encoded in the sequence order of nucleotides in the DNA sequence and the human genome (*Homo sapiens*) comprises of about 3 billion nucleotides present within the 23 human chromosomes (Lander et al., 2001; Venter et al., 2001).

The nucleotide is the basic and underlying structural unit of DNA and the arrangement and order of the four nucleotides in DNA is also the functional basis of storing and transmitting information. Knowledge and understanding of the sequence of nucleotides in DNA is a fundamental paradigm for molecular medicine. The central role of DNA for genetic information storage has increased immensely the importance of DNA sequencing. Furthermore, DNA sequencing techniques are key tools in many scientific fields and a considerable number of different sciences have been benefiting from these techniques, ranging from molecular biology, genetics, biotechnology, forensics to archeology and anthropology. DNA sequencing is also promoting new discoveries that are revolutionizing the conceptual foundations of many fields. It is also considered to be the golden standard method for accurate identification of microorganisms.

The chain termination sequencing method, also known as Sanger sequencing, developed by Frederick Sanger and colleagues (Sanger et al., 1977), has been the most widely used DNA sequencing method since its advent in 1977 and still is in use after more than 26 years since its development. The remarkable advances in chemistry and automation to the Sanger sequencing method has made it to a simple and elegant
technique, central to almost all past and current genome-sequencing projects of any significant scale. Despite all these grand advantages, there are limitations in this method, which could be complemented with other techniques.

Pyrosequencing technology is a rather novel DNA sequencing technology, developed at the Royal Institute of Technology (KTH), and is the first alternative to the conventional Sanger method for de novo DNA sequencing. This method relies on the luminometric detection of pyrophosphate that is released during primer-directed DNA polymerase catalyzed nucleotide incorporation. At present, it is suited for DNA sequencing of up to one hundred bases and it offers a number of unique advantages.

In this thesis the main DNA sequencing technologies will be briefly reviewed with the emphasis on the Pyrosequencing technology. The central focal point of this thesis will be on methodological and chemical advancements of the Pyrosequencing technology and its applications and troubleshooting in microbial and viral typing.
2. Polymerase chain reaction

PCR is an acronym, standing for polymerase chain reaction, which is a versatile technology that is advancing disease detection, and is one of the revolutionizing scientific developments. Kary Mullis developed the PCR technique in 1983 and it has been methodically improved since then (Mullis and Faloona, 1987). This innovation resulted in a Nobel Prize in chemistry in 1993 for Kari Mullis.

PCR is principally a primer extension reaction for amplifying specific nucleic acids in vitro. The use of PCR in molecular diagnostics has increased to the point where it is now adopted as a standard tool for detecting nucleic acids from a number of origins and it has become a key tool in research and diagnostics.

The use of a thermostable polymerase (most commonly derived from the thermophilic bacterium *Thermus aquaticus* and called *Taq*) has allowed the separation of newly produced complimentary DNA and successive annealing or hybridization of primers to the target DNA with minimal loss of enzymatic activity. Exponentially, PCR permits a short part of DNA to be amplified to about a billion fold, which in turn, allows determination of size, nucleotide sequence, etc. Since the advent of PCR, measurable improvements and modifications have been achieved in this technique, such as multiplex PCR (Chamberlain et al., 1988), asymmetric PCR (Shyamala and Ames, 1989), hot-start PCR (Chou et al., 1992), nested PCR (Haqqi et al., 1988), RT-PCR (Erlich et al., 1991), touchdown PCR (Don et al., 1991), real-time PCR (Holland et al., 1991; Gibson et al., 1996; Heid et al., 1996), and miniaturization (Wittwer et al., 1997; Kopp et al., 1998).

In principle, a pair of synthetic oligonucleotides or primers with each hybridizing to one strand of a double-stranded DNA target are used in the PCR reaction. The hybridized primers perform as substrate for the DNA polymerase, which builds a complementary strand via sequential incorporation of deoxynucleotides. The process can be summarized in three steps: (i) dsDNA separation at temperatures >90°C, (ii) primer annealing at 50-75°C, and (iii) optimal extension at 72-78°C (Mackay et al., 2002). The rate of temperature alterations, or in other words, ramp rate, the time-span of the incubation at each temperature and the number of times each set of temperatures (or cycle) repeated are controlled by a thermocycler. The advance of
technologies in the current years has considerably decreased the ramp times utilizing electronically controlled heating blocks or fan-forced heated air flows to regulate the reaction temperature.
3. DNA sequencing techniques

Recent efforts to blueprint the entire genome of organisms ranging from yeast to humans have yielded more high-throughput methods of DNA sequencing. Despite the increased demand for these high-throughput advanced projects, the standard technique for DNA sequencing is still the one developed in 1970s by Sanger and colleagues. This invention resulted in that Frederick Sanger received the Nobel Prize in Chemistry for this elegant technique in 1980. The Sanger dideoxy technique is based on the electrophoretic separation and detection of synthesized single-stranded DNA molecules that have been terminated with dideoxynucleotides. Another concurrent sequencing strategy, known as Maxam-Gilbert sequencing, utilizing chemical degradation was described also in 1977. This method was not widely adopted due to the use of toxic chemicals (Maxam and Gilbert, 1977).

In this chapter, the focus will be aimed at de novo sequencing methods and technical parameters, such as sequencing speed, read length, and base-call precision.

3.1. Dideoxy DNA sequencing by chain termination

The Sanger DNA sequencing technique (Sanger et al., 1977) has revolutionized the biological sciences. As mentioned earlier, this simple and elegant method is still in use after a quarter of century with continual advancements in chemistry and automation. The principle behind the Sanger method is the termination of the enzymatically synthesized DNA. The dideoxynucleotide (ddNTP) is incapable of forming the next bond in the DNA chain and consequently, synthesis of the DNA sequence is interrupted when a ddNTP is incorporated in the growing chain. The four reactions with the labeled ddNTPs are separated electrophoretically based on their molecular weight, and the different-sized DNA fragments are read from smallest to largest.

The need for a DNA sequencing technique to have the possibility to provide long read-length (number of bases read per run), short analysis time, low cost, and high accuracy has brought about establishment of several modifications of the original Sanger dideoxy technique. Automation has been a major aspect of these
improvements as it has reduced the time and resources required for sequencing of long stretches of DNA.

Cycle sequencing is a modification of the traditional Sanger sequencing method. In principle, cycle sequencing is exactly the same as Sanger sequencing, but the reaction proceeds in a series of cycles (Innis et al., 1988; Carothers et al., 1989; Murray, 1989). The discovery of the PCR and the use of the heat stable Taq polymerase facilitated the ability to perform sequencing with reduced amounts of DNA template as the sequencing reaction can be repeated over and over again in the same tube. Thus, less DNA template is needed than for conventional sequencing reactions. The DNA template is denatured by heating, followed by primer annealing and extension (incorporation of ddNTPs) using a thermostable DNA polymerase. The cycle is then repeated. In this way, several dideoxy-terminated chains are synthesized from one template strand. The thermal denaturation step is carried out in the presence of primers, preventing the re-annealing of double-stranded template. The large amount of product produced from a single template strand means that this technique is far more sensitive than standard Sanger sequencing.

Many enzymes are available for PCR and cycle sequencing. Some PCR enzymes have an extra feature, a 3'-5' exonuclease activity. This feature is called the proof-reading ability of the enzyme, i.e. its ability to correct mistakes made during incorporation of the nucleotides. For cycle-sequencing this activity must be suppressed to avoid un-interpretable data.

3.1.1. Slab gel electrophoresis

Acrylamide slab gel electrophoresis has been the most widely adopted format for Sanger sequencing with ABI PRISM and ALFexpress as two of the broadly used sequencers. The principle of the instruments is based on detection of fragment bands as they drift in the gel by a scanning fluorescent detection system. Read length of 650-750 bases can be expected from these instruments, which is still a widely used sequencing platform despite suffering from the common drawbacks of slab gel instruments: gel casting and lane tracking.
3.1.2. Capillary electrophoresis

Relying merely on slab gel technology was not sufficient to accomplish the challenges set by the Human Genome Project. The continual and unrelenting drawbacks of slab gel electrophoresis and the demand for more rapid sequencing and higher throughput led to the development of capillary electrophoresis (CE) (Huang et al., 1992). The completion of the human genome was only possible due to several technological advances offered by CE.

Capillary electrophoresis is a fast technique for separation and analysis of biopolymers. The high surface-to-volume ratio of a capillary allows more rapid heat dissipation than is possible in slab gels, therefore allowing higher operating voltages, and consequently, faster sequencing reactions could be achieved. This higher electric field and faster separation has made CE approximately 8-10 times faster than conventional slab gel electrophoresis. Electrophoresis is performed basically in an approach similar to slab gels with the advantage that each capillary contains a single DNA sample and therefore tracking problems are eliminated.

3.2. Non-electrophoretic DNA sequencing methods

A few DNA sequencing methods which are not electrophoretic based have been conceived in the past two decades. These techniques have advantages and disadvantages compared to electrophoretic separation methods, depending on the type of applications performed.

3.2.1. Mass spectrometry

Significant improvements in the ionization of biopolymers in the gas phase has made mass spectrometry (MS) an alternative for fast and accurate DNA sequencing, particularly with the approach Matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF-MS), which was first introduced by Karas and Hillenkamp (Hillenkamp et al., 1991). This method allows the macromolecule to be desorbed as an intact gas-phase ion by embedding it in the crystal of a low-molecular-weight molecule that strongly absorbs energy from a pulse of laser light. In MALDI-TOF-MS the Sanger DNA molecule are desorbed and ionized and then subjected to an intense electric field, accelerating the fragments in a flight tube in
common kinetic energy, which is relative to their mass-charge-ratio. The detection is based on time required for each particle to collide with an ion-to-electron conversion detector.

In essence, MALDI-TOF-MS DNA sequencing adopted the enzymology and nucleic acid chemistry established for conventional sequencing. MALDI-TOF-MS read length is stated to be limited by strand fragmentation (Smith, 1996) and salt-adduct formation (Lin et al., 1999), and it relies on strand stringent purification. In contrast, there is currently no prospect of reaching long-reads per reaction, as is common with high-throughput capillary devices. Therefore, MALDI-TOF-MS is unlikely to be widely used in de novo genomic sequencing in the foreseeable future.

3.2.2. Sequencing-by-hybridization

The principle underlying the sequencing-by-hybridization (SBH) technique is in actual fact the same as Southern blotting (Southern, 1975). Two research groups independently described the notion of SBH in 1988 (Lysov Iu et al., 1988; Drmanac et al., 1989). De novo sequencing was one of the major foundations for developing hybridization arrays (Drmanac et al., 1992; Drmanac et al., 1993). SBH is based on annealing a labeled unknown DNA fragment to a complete array of short oligonucleotides (e.g. all 65,536 combinations of 8-mers) and analyzing the unidentified sequence from the hybridization pattern by computer-assisted assembling programs, reconstructing the sequence order of the DNA fragment.

Nevertheless, SBH has faced problems of ambiguous sequence results connected with repetitive regions within the unknown DNA sequence, and in addition, formation of secondary structures in the target DNA is an obstacle for this method. Moreover, small variations in duplex stability between a complete match and a single mismatch (false positive), which can happen at the 3’-terminus, is another limitation. These problems can be addressed in expression analysis or comparative sequencing on microarray format by SBH, but they have not been addressed in de novo sequencing yet.

3.2.3. Sequencing-by-synthesis

The sequencing-by synthesis (SBS) approach was first introduced by Robert Melamede in 1985 (Melamede, 1985). The principle of SBS is based on sequential
nucleotide addition and incorporation in a primer-directed polymerase extension reaction in a way that a complementary strand will be formed by the iterative addition of the four dNTPs. Figure 1 illustrates the principle of SBH.

The event of nucleotide incorporation in the primed DNA can be detected directly or indirectly. In the direct method, dNTPs are fluorescently labeled and the detection is performed by a fluorometric detection instrument (Canard and Sarfati, 1994; Metzker et al., 1994). This approach is limited to only a few bases as the nucleotide incorporation efficiency is low, which consequently can lead to unsynchronization (Metzker et al., 1994). In addition, removal of the fluorophore from the incorporated nucleotides in each cycle is another extra step for this approach.

In the indirect detection approach, natural nucleotides are utilized in the incorporation process. The nucleotide incorporation results in release and detection of pyrophosphate (PPi) (Nyren, 1987; Hyman, 1988). The major advantage of this approach is utilization of non-modified nucleotides, and therefore, getting the best of polymerase incorporation efficiency in the assay. The enzymatic process shown beneath is established on luminometric detection of ATP based on conversion of PPi through a cascade of enzymatic reactions.

\[(\text{DNA})_n + \text{nucleotide} \xrightarrow{\text{polymerase}} (\text{DNA})_{n+1} + \text{PPi}\]

\[\text{PPi} \xrightarrow{\text{ATP sulfurylase}} \text{ATP}\]

\[\text{ATP} \xrightarrow{\text{luciferase}} \text{Light}\]
Figure 1. Schematic illustration of the sequencing-by-synthesis principle, in which a primed DNA template is subjected to iterative nucleotide additions in the presence of DNA polymerase (the oval shape). If the added nucleotide is complementary, it will be incorporated by the polymerase resulting in release of PPi and extension of the DNA template.
When dNTP is added to the reaction mixture, DNA polymerase incorporates the complementary nucleotide onto the 3' terminus of the growing strand, resulting in the release of PPI. The liberated PPI in the system is monitored using a coupled enzymatic reaction in which ATP sulfurylase converts PPI to ATP. The produced ATP is sensed by the enzyme firefly luciferase, which in turn leads to generation of light (Nyren, 1987; Nyren et al., 1993). A photon multiplier or a charge coupled device (CCD) camera then records the visible light. Hyman studied the enzymatic SBS for DNA sequencing by using a gel-filled column to attach nucleic acid template and the DNA polymerase while solutions containing the four dNTPs were pumped through one at a time. The generated PPI was then measured off-line by a device consisting of a series of columns containing covalently attached enzymes. This method is not very suitable for robust DNA sequencing and no further progress of this study is reported in the literature.

Progressively in time, the sequencing-by-synthesis approach was developed to a user-friendly and robust DNA sequencing method by more efficient removal of the nucleotides and further modifications of the sequencing-by-synthesis principle, resulting in innovation of Pyrosequencing technology, which was developed by Nyrén and coworkers (Ronaghi et al., 1996; Nyren et al., 1997; Ronaghi et al., 1998b; Nyrén, 2001).
4. Pyrosequencing technology

Removal or degradation of unincorporated or excess dNTPs was a crucial factor for sequencing-by-synthesis in order to be applied for DNA sequencing. In Pyrosequencing nucleotide removal is performed in two different ways: (i) the solid-phase Pyrosequencing, which utilizes a three-coupled enzymatic procedure with washing steps and (ii) the liquid-phase Pyrosequencing technique, which employs a cascade of four enzymes with no washing steps.

4.1. Pyrosequencing by the solid-phase approach

The solid-phase method is based on a combination of the sequencing-by-synthesis technique and a solid-phase technique. The four nucleotides are dispensed sequentially in the reaction system and a washing step removes the unincorporated nucleotides after each addition. There are different immobilization techniques that can be used for the solid-phase approach. In one approach, the biotin-labeled DNA template with annealed primer is immobilized to streptavidin coated magnetic beads (Ronaghi et al., 1996). The immobilized primed single stranded DNA is incubated with three enzymes: DNA polymerase, ATP sulfurylase and luciferase. After each nucleotide addition to the reaction mixture, the DNA template is immobilized by a magnet system and the unincorporated nucleotides are removed by a washing step.

Loss of DNA templates in the washing procedure, repetitive addition of enzymes, unstable baseline fluctuations and automation difficulties are drawbacks of this approach. Another approach, which seems more promising, is the biotin-mediated immobilization of ATP sulfurylase and luciferase on streptavidin coated nylon tubes for detection of PPi and ATP by using a flow system. This may result in a more stable baseline and improved sequence signals. Furthermore, the later approach is more cost-effective as replenishment of the enzyme mixture after each dNTP addition is eliminated and has the advantage that inhibitory products are not accumulated during the process. In addition, it is compatible with microfluidic formats with the possibility for miniaturization.

The solid-phase Pyrosequencing is being applied in microfluidic format by 454 Life Sciences. By applying this format, DNA templates are immobilized on a solid
support and it is also possible to immobilize the detection enzymes (ATP sulfurylase and luciferase) onto a solid support to reduce reagent cost. In an attempt the adenovirus genome has been sequenced by utilizing the solid-phase Pyrosequencing principle, which is the first demonstration of sequencing an entire genome. The company claims to be making significant progress towards the goal of whole genome sequencing in a massively parallel fashion. The 454’s developed microfluidic instrument sequences DNA within thousands of 75 pico-liter wells contained within the specifically designed PicoTiter Plate. The high density PicoTiter plate allows amplification and sequencing of several thousand to several hundred thousand DNA samples simultaneously. Figure 2 illustrates the microfluidic system in the PicoTiter plate.

![Figure 2](image_url)

**Figure 2.** Schematic demonstration of the 454 fluidic system. The PicoTiter plate is placed face down on the backplate, creating a sealed, closed flow chamber. Sequencing reagents flow through the fluidic system in a programmed order, and the results are read by a detector. This figure is printed by the permission of 454 Life Sciences.

DNA fragments are first amplified within the wells, eliminating the need for off-line amplification, which is less cost and time-effective. The reagents can diffuse uniformly into and out of the wells. The light generated by the nucleotide
incorporations in the system, is detected by a light detector. Figure 3 shows a close-up of the fluidic system in the wells and figure 4 illustrates the overall flow and detection system.

**Figure 3.** Schematic illustration of flow in the solid-phase Pyrosequencing method. Each well contains 75 picoliter volume where the amplification and sequencing is performed. The enzymes and DNA template are immobilized on the beads. This figure is used by the permission of 454 Life Sciences.

**Figure 4.** Schematic illustration of the overall flow and detection system, demonstrating flow of reagents and nucleotides in the microfluidic flow system and detection of light by a camera. This picture is adopted from 454 Life Sciences.
The microfluidic format for the solid-phase Pyrosequencing method exhibits promising applications for large whole genome sequencing projects.

In short, performing Pyrosequencing reaction in a microfluidic system offers several advantages: (i) theoretically, sequencing can be performed faster since the cycling time for each nucleotide addition can be reduced, (ii) as aforementioned, accumulation of inhibitory substances will be minimized since washing is performed after each nucleotide addition/cycle, (iii) low amounts of enzymes and reagents will be consumed and (iv) the possibility for parallel processing of numerous samples simultaneously.

4.2. Pyrosequencing by the liquid-phase approach

The breakthrough for Pyrosequencing by the liquid-phase approach came about by the introduction of a nucleotide-degrading enzyme, called apyrase (Nyren, 1994; Ronaghi et al., 1998b; Nyrén, 2001). The implementation of this enzyme in the Pyrosequencing system excluded the use of solid-phase separation, and consequently, eliminated extra steps such as washes and repetitive enzyme additions. The apyrase shows high catalytic activity and low amounts of this enzyme in the Pyrosequencing reaction system efficiently degrade the unincorporated nucleoside triphosphates to nucleoside diphosphates and subsequently to nucleoside monophosphate. In addition, apyrase stabilizes the baseline with no fluctuations in the sequencing procedure as the same enzymes catalyze the reaction continuously.

The liquid-phase Pyrosequencing method employs a cascade of four enzymes and the DNA sequencing is monitored in real-time. The sequencing reaction is initiated by annealing a sequencing primer to a single-stranded DNA template. Figure 5 shows schematically in detail sequencing of a partially amplified DNA of human papillomavirus (HPV) by Pyrosequencing technology. The cascade of the four-enzyme catalyzed reactions is demonstrated where APS stands for adenosine 5'-phosphosulphate and $h\square$ represents light photons emitted by the bioluminoscent reaction.
Figure 5. The liquid-phase Pyrosequencing method is a non-electrophoretic real-time DNA sequencing method that uses the luciferase-luciferin light release as the detection signal for nucleotide incorporation into target DNA. The four different nucleotides are added iteratively to a four-enzyme mixture. The pyrophosphate (PPi) released in the DNA polymerase-catalyzed reaction is quantitatively converted to ATP by ATP sulfurylase, which provides the energy to firefly luciferase to oxidize luciferin and generate light (hν). The light is detected by a photon detection device and monitored in real time by a computer program. Finally, apyrase catalyzes degradation of nucleotides that are not incorporated and the system will be ready for the next nucleotide addition. The sequence presented here identifies the genotype of a human papillomavirus (HPV-18).
In the liquid-phase Pyrosequencing method, the sequencing primer is hybridized to a single-stranded DNA template and mixed with the enzymes, DNA polymerase, ATP sulfurylase, luciferase and apyrase, and the substrates APS and luciferin. The four dNTPs are added to the reaction mixture sequentially. When the DNA polymerase incorporates the complementary nucleotide onto the 3’-end of the primed DNA template, PPI is released in a quantity equimolar to nucleotide incorporation. ATP sulfurylase quantitatively converts PPI to ATP in the presence of APS. The produced ATP drives the luciferase-mediated conversion of luciferin to oxyluciferin that generates visible light in amounts that is proportional to the amount of ATP. The light produced in the luciferase-catalyzed reaction is detected by a photon multiplier tube or CCD camera and are recorded as a peak in a pyrogram. Each light signal is proportional to the number of nucleotides incorporated. Apyrase continuously degrades ATP and unincorporated excess dNTPs. When degradation is complete, another dNTP is added. Addition of dNTPs is performed one at a time. As the process continues, a complementary DNA strand is formed and the nucleotide sequence is determined from the signal peaks in the pyrogram.

Pyrogram in Pyrosequencing method corresponds to electropherogram in Sanger DNA sequencing. A pyrogram can be observed in real-time while the sequencing reaction is carried out and it is virtually the display of all the sequence signal peaks, showing the nucleotide additions, the order of nucleotide dispensation and dNTP incorporations and non-incorporations, which consequently results in base calling.

Pyrosequencing technology takes advantage of the co-operativity of several enzymes to monitor DNA synthesis. Parameters such as stability, fidelity, specificity, sensitivity, $K_M$, and $k_{cat}$ are imperative for the optimal performance of the enzymes used in the sequencing reaction. The kinetics of the enzymes can be studied in real-time as shown in figure 6. The slope of the ascending curve in a pyrogram displays the activities of DNA polymerase and ATP sulfurylase, the height of the signal shows the activity of luciferase, and the slope of the descending curve demonstrates the nucleotide degradation.
Figure 6. Interpretation of a sequence signal in the pyrogram. The details of this figure is described in the text.

As mentioned, apyrase catalyzes degradation of ATP, unincorporated and excess nucleotides to yield nucleotide monophosphates and inorganic phosphate, resulting in a steady descending of the light output to baseline in the pyrogram. When the baseline is achieved, there will be a time interval of less than 1 minute for the nucleotides to be fully degraded and then the next dNTP can be added. The nucleotide addition cycle can be repeated numerous times without involvement of washing steps.

The enzymes in the Pyrosequencing method originate from different organisms. The standard Pyrosequencing chemistry utilizes the modified 3’-5’exonuclease-deficient Klenow fragment of *E. coli* DNA polymerase I, which is a relatively slow polymerase (Benkovic and Cameron, 1995). The ATP sulfurylase used in the Pyrosequencing method is a recombinant version from the yeast *Saccharomyces cerevisiae* (Karamohamed et al., 1999) and the luciferase is from the American firefly *Photinus pyralis*. The apyrase is from *Solanum tuberosum* (Pimpernel varieity) (Espinosa et al., 2003; Nourizad et al., 2003). The overall time frame from polymerization to light detection takes place within 3-4 seconds at room temperature. ATP sulfurylase converts PPi to ATP in approximately 1.5 seconds and the generation of light by luciferase takes place in less than 0.2 seconds (Nyren and Lundin, 1985). The generated light with a wavelength maximum of 560 nanometers can be detected by a photodiode, photon multiplier tube, or a CCD-camera.
The Pyrosequencing technology has many unique advantages among DNA sequencing technologies. One advantage is that the order of nucleotide dispensation can be easily programmed and alterations in the pyrogram pattern reveal mutations, deletions and insertions (this will be discussed in the next chapter). Another major advantage of the Pyrosequencing method is that sequencing is performed directly downstream of the primer, starting with the first base after the annealed primer, making primer design more flexible. And moreover, the Pyrosequencing technique is carried out in real-time, as nucleotide incorporations and base callings can be observed continuously for each sample. In addition, the Pyrosequencing method can be automated for large-scale screenings.

4.3. Template preparation for the Pyrosequencing method

PCR products that are to be sequenced contain excess amounts of PCR primers, nucleotides and PPI, which have to be removed or degraded prior to sequencing by the Pyrosequencing method. These by-products can interfere in the course of DNA sequencing since the amplification primers can re-anneal to the amplified DNA, causing unspecific sequence signals, and the remaining nucleotides and PPI in the PCR products should also be removed or degraded before the sequencing procedure is carried out in order to eliminate the risk for asynchronous extension and depletion of APS. The salt in the PCR reaction slightly inhibits the enzyme system and should be removed or diluted.

The use of an internal primer for DNA sequencing is preferential for higher specificity. Since there is the possibility for unspecific amplification products in PCR, especially for genomic DNA, utilizing internal primers for sequencing can improve the sequence signal quality.

There are two approaches currently used for template preparation: single-stranded template preparation and double-stranded template preparation.

4.3.1. Single-strand approach

As the name suggests, this approach is based on sequencing of single-stranded DNA template. In traditional Pyrosequencing method, solid-phase purification of the template, based on streptavidin biotin binding has been utilized. One of the PCR
primers is biotin-labeled at the 5'-terminus. The biotinylated PCR product binds to a streptavidin coated solid support, allowing the immobilization of the DNA fragment. This permits the immobilized DNA to be single-stranded by alkali treatment (high pH), and simultaneously, the interfering components to be removed. The eluted non-biotin-labeled strand can also be collected and used for sequencing. After the separation step, a sequencing primer is annealed to the single-stranded DNA. There are two methods available for single-strand DNA separation: (i) streptavidin coated magnetic beads, using a magnet for purification, denaturation and elution, and (ii) streptavidin coated Sepharose beads, based on filter separation by a vacuum system. Figure 7 shows the magnetic separation approach. An automated magnetic separation system is available that is capable of performing this procedure on 96-well plate format. The automated system can be programmed for different procedures, such as extra washing step to remove excess sequencing primers prior to DNA sequencing, and regeneration of magnetic beads for multiple use.

4.3.2. Double-strand approach

There are two different methods available for preparation of double-stranded template. In the first method (Nordstrom et al., 2000a), nucleotides, PPi and amplification primers are removed enzymatically. Nucleotides and PPi are degraded by addition of apyrase and pyrophosphatase, and the amplification primers are removed by exonuclease I, which specifically degrades single-stranded DNA. The double-stranded DNA is then denatured by heat and the sequencing primer is annealed.

The second method is based on using blocking primers (Nordstrom et al., 2002). The principle behind this approach is to prevent the re-annealing of the present PCR primers to the amplified DNA. This could be carried out either by utilizing complementary blocking primers hybridizing to the PCR primers or using non-extendable primers, with the same sequence as the PCR primers, binding to the amplification sites. This later method is faster and the sequence results obtained show better signal quality.
Figure 7. Single-stranded template preparation prior to the Pyrosequencing method utilizing steptavidin conjugated magnetic beads. The biotinylated amplified DNA binds to the magnetic beads and is separated under alkali treatment. The sequencing primers can be annealed to either strand for sequencing.
5. Present investigation

The present investigation deals with method improvements of the Pyrosequencing technology, which have led to increased read-length, enhanced sequence performance and sequencing of challenging homopolymeric regions. A novel approach for typing of clinical specimens harboring multiple infections (a multitude of different types/species) as well as unspecific amplification products from genomic DNA will be discussed. In addition, it will also be shown that these advances in Pyrosequencing technology have contributed to accomplishment of new significant applications.

5.1. Methodological improvements of Pyrosequencing technology

Pyrosequencing at its advent faced limitations for de novo sequencing of about 20 nucleotides downstream of the annealed primer. This limitation was due to accumulation of by-products in the system, giving rise to increased background and lower sequence-specific signals. Improvements of the Pyrosequencing chemistry for longer DNA-reads and enhancement of sequencing quality in general and for regions containing homopolymeric regions have resulted in substantial improved sequence data. Furthermore, these methodological advancements have contributed to utilization of the Pyrosequencing technique for different microbial and viral applications.

5.1.1. Read-length improvement (Paper I)

Despite the advantages of Pyrosequencing technology such as accuracy, flexibility, parallel processing and use of non-labeled primers and nucleotides, it was merely being used for short reads. In many cases, longer reads were not possible due to the continuous decrease of nucleotide degrading efficiency of apyrase during the sequencing procedure. The possibility to improve the method for longer reads should open new avenues for new applications in many different areas.

During the development of the Pyrosequencing method, the natural dATP was replaced by dATP[S] to increase the signal-to-noise ratio (Ronaghi et al., 1996). The natural dATP is a substrate for luciferase giving simply rise to false-positive signals. However, the performance of the method for longer reads was still limited by this substitute. We observed continuous decrease of the apyrase activity during the
sequencing process and inefficient incorporation of nucleotide A in homopolymeric T-regions. The applied dATP\(\alpha\)S was a mixture of two isomers (Sp and Rp) (Paper I). The structures of the two isomers are shown in figure 8.

Figure 8. Structures of the Sp and Rp isomers of dATP-\(\alpha\)-S

Different concentrations of Sp/Rp-mix and pure Sp-isomer were comparatively analyzed on ATP-luciferase assays, and furthermore, on different DNA templates. The experimental results showed that higher nucleotide concentration of the Sp/Rp-mix was needed for efficient incorporation of nucleotide A in homopolymeric regions. In addition, pure Sp-isomer could be applied for the same nucleotide A incorporation experiments by only half the concentration. We also showed that the Rp-isomer was not a substrate for the Klenow polymerase and this was also in accordance with earlier studies (Burgers and Eckstein, 1979). Significant sequence results were achieved for longer DNA sequencing by the new pure Sp-nucleotide. The observed inhibition of the apyrase activity is probably due to accumulation of degradation products, mainly nucleoside diphosphates, as no inhibition was observed in the presence of alkaline phosphatase. Our experiments also indicate that the diphosphates of the alpha-S nucleotides are stronger inhibitors than the diphosphates of the natural nucleotides, and that increased concentration of dATP\(\alpha\)S nucleotides had negative effect on the
apyrase activity. The apyrase activity continuously decreased during the successive addition of the nucleotides.

Furthermore, these investigations clearly showed impressive improvements in the read length and the sequencing performance when using lower nucleotide concentrations by using the pure Sp-isomer. The pure Sp-dATP[S] was utilized for sequencing of numerous DNA templates. Read lengths surpassing 100 bases were achieved although it is worth noting that the read length in Pyrosequencing technology is template dependent. This will be discussed further in this chapter.

5.1.2. Advancements in sequencing T-homopolymeric regions (Paper II)

Despite the increase in the read length by the pure Sp-dATP[S], there are templates that cannot be sequenced synchronously due to homopolymeric T-regions. The Klenow polymerase normally used in the Pyrosequencing method shows low catalytic efficiency for the modified alpha thio nucleotides. This has yielded in incomplete incorporation of dATP[S] in templates containing homopolymeric regions with more than 4-5 T-nucleotides, resulting in non-synchronized extensions.

To obtain high quality data with the Pyrosequencing technology, the DNA polymerase should fulfill a number of characteristics that are essential for DNA sequencing. The DNA polymerase must be 3’-5’exonuclease-deficient in order to avoid degradation of sequencing primer annealed to the template, and incorporate nucleotides with good fidelity and with reasonable processivity. Moreover, the DNA polymerase must be able to utilize dATP analogs (such as dATP[S]) and to incorporate such nucleotides with high efficiency across homopolymeric regions. Furthermore, it is advantageous if primer-dimers and loop-structures are not be utilized as substrates for the polymerase. Traditionally, the exonuclease-deficient Klenow DNA polymerase has been used in the Pyrosequencing chemistry since it displays many of these important features. However, despite the fact that Klenow polymerase fulfills many of these characteristics sequencing problems may still be encountered due to inefficient incorporation of nucleotide dATP[S] in homopolymeric T-regions and extension of mispairs formed by unspecific hybridization events.
**Figure 9.** Sequencing performance by two different DNA polymerases. The Pyrosequencing reaction was performed on a PCR fragment, containing a poly-T track of seven T, in the presence of (a) Klenow, and (b) Sequenase. Note the low T and high A signal peaks obtained (filled and dotted arrows, respectively) when Klenow was applied. When Sequenase was used the sequence could be easily read and the T and A signal peaks were close to the expected levels. The order of nucleotide addition is indicated on the bottom of the traces. The sequence obtained is indicated above trace b.

In order to improve nucleotide incorporation efficiency, a modified T7 polymerase (exonuclease deficient T7 DNA polymerase), Sequenase™, was selected based on known characteristics. We evaluated the effect of Sequenase on the Pyrosequencing performance on poly(T)-rich templates in parallel with the Klenow DNA polymerase. In all combinations tested, Sequenase demonstrated far better catalytic efficiency for the homopolymeric T-regions. On the other hand, Klenow
polymerase resulted in unsynchronized sequence signals in the parallel sequencing reactions on templates containing regions with 4 to 5 T. Sequenase has the ability to incorporate dATP’s more efficiently and we observed synchronized extension in templates containing regions with 7 to 8 T as demonstrated in figure 9.

Our experimental results both Klenow and Sequenase incorporate all the other nucleotides with good efficiency.

An additional advantage with Sequenase was the ability to discriminate between specific and unspecific binding of the sequencing primer to the template and to itself (primer-dimer). In contrast to Klenow polymerase, Sequenase was also reluctant to extend free 3’-ends formed by loop-structures. It has been indicated that Sequenase requires a longer stem (15 bases) than Klenow polymerase (12 bases) if a loop-structure was used as primer (Ronaghi et al., 1998a). It has also been shown that Sequenase is more discriminating for 3’-mismatches than Klenow polymerase (Nyren et al., 1997).

Our experiments show that the use of Sequenase instead of Klenow polymerase in the Pyrosequencing method has several advantages, such as (i) more efficient polymerization in homopolymeric T-regions, (ii) longer reads on difficult poly-T rich templates, (iii) better sequencing performance in the presence of primer-dimers and self-priming templates. Therefore, we predict that the use of Sequenase in the Pyrosequencing method might lead to new applications for this technology in the field of long read DNA sequencing for challenging templates.

5.1.3. Improved DNA sequencing by SSB (Paper VIII)

A protein, which has shown an essential role in sequence quality of Pyrosequencing technology, is single-stranded DNA-binding protein (SSB). Introduction of SSB to the Pyrosequencing chemistry has facilitated the optimization of different parameters (Ronaghi, 2000; Ehn et al., 2002). SSB has improved long read DNA sequencing, sequencing of challenging templates, making primer design more flexible and eliminating the need for intermediate washes to remove primers from the reaction mixture due to problems such as primer-dimers and mis-hybridization of primers.
Figure 10 demonstrates two HPV amplicons sequenced both in the presence and in the absence of SSB (with an extra wash performed prior to the Pyrosequencing reaction to remove excess primers); the sequence results for HPV-72 sequenced by the GP5+ primer and HPV-31 (simulated amplicon mixture of HPV-31/40/73) sequenced by a seven-primer-pool are remarkably improved by the presence of SSB as shown. Consequently, SSB was used in all further Pyrosequencing reactions.

Figure 10. Effects of SSB on improvement of sequence signal quality. (a) shows pyrograms from sequencing of HPV-72 by the GP5+ primer in the absence and in the presence of SSB. (b) shows sequencing of HPV-31 in a simulated multiple infection of HPV-31/40/73 sequenced by a seven-primer pool both in the absence and in the presence of SSB.
5.1.4. Pre-programmed DNA sequencing (*Paper III*)

Pyrosequencing technology has the unique and significant advantage of a pre-programmed nucleotide addition strategy for different purposes. This is considered as an important benefit, since nucleotides are dispensed into the reaction mixture in a programmed order. In contrast, *de novo* sequencing uses cyclic nucleotide dispensation, which can result in longer reaction time, together with product accumulation due to more nucleotide dispensations in the system. The pre-programmed feature can be used as a tool for mutation/artifact detection (Garcia et al., 2000).

In our work, we have used this approach for clone checking. Given that DNA manipulation techniques can result in artifacts, methods that facilitate quality control are important. In the case of cloning and/or mutagenesis of a gene or specific DNA fragment, the new construct must be carefully checked for proper function. Clone checking is necessary to confirm that a DNA fragment has been inserted in the correct orientation and frame within a vector. Moreover, it is essential to check for undesired mutations that could have occurred during initial PCR steps. In addition, the quality of oligonucleotides utilized during different procedures must be considered.

We have used the pre-programmed dispensation ability of Pyrosequencing technique for such quality control with successful results. Mutations and deletions could be easily observed in the pyrogram sequence pattern as missing signal peaks. It is worth mentioning that single-base mutations, insertions or deletions that occur in a context of similar bases, are observed as alterations in signal peak heights within a homopolymeric region. We used this approach for mutation/deletion/insertion detection for clone checking of the apyrase gene by sequencing a pre-programmed sequence pattern of up to 150 bases. Figure 11 shows three analyzed colonies: (i) with no mutation in the N-terminus region (figure 11a), (ii) with one base deletion in nucleotide A (figure 11b) and (iii) with a large deletion of 195 nucleotides (figure 11c). It is worth noting that although a pre-programmed nucleotide dispensation strategy was used, the sequence after the large deletion in figure 11c, could be identified. Thus, after the large deletion the following sequence was obtained: CCTATTGGCAACAAATATTGAGTTTTAT. This sequence is located 123 bases into the apyrase gene.
The pre-programmed dispensation resulted in a more efficient clone checking as longer and faster reads, and more time effective detection of artifacts was achieved. For longer fragments, primers that bind to different regions of a DNA fragment, separated by 100-150 bases, could be used to address the read length limitation.

**Figure 11.** Pyrogram, pre-programmed sequence information, from three of the apyrase clones analyzed. (a) one of the correct clones, (b) one of the clones with a one base deletion (the deletion is indicated with an arrow), and (c) a clone with a large deletion (the start position for the sequence after the deletion is indicated with an arrow). The correct sequence of a part of the region analyzed is shown above trace (a). The nucleotides were added according to the correct sequence of the target DNA.

5.1.5. Multiple sequencing-primer method and applications (Papers IV and V)

DNA sequencing is considered as the gold standard method for accurate typing and identification of microorganisms and viruses. DNA sequencing is based on nucleic acid sequence knowledge; the informational core of every organism. Besides, it is more reliable than the hybridization-based techniques, which are based on signal detection and discrimination of closely related species. The hybridization techniques are sensitive to possible cross-hybridization of the probes, and in addition, false positive results are unlikely to be distinguished. In contrast, one can draw clearly this distinction in DNA sequencing as it provides nucleic acid knowledge.
In microbial and viral typing, universal/consensus and degenerate primers that bind to conservative regions are often used for amplification of microorganisms and viruses (Woese, 1987; Larsen et al., 1993; Maidak et al., 1994). By utilizing general primers, a broader range of species/types can be covered in the amplification process. Accordingly, the variable regions on the microbial amplified DNA are used for detection and identification. The variable region is a stretch of nucleotides, which shows varying sequence between different nucleic acid molecules covering the same gene. The microbial amplicons may contain more than one species or type (we will apply the term multiple infections for different multitude of species or types). In such cases, all the present types or species will be subjected to amplification when utilizing consensus, general or degenerate primers. In addition, consensus and degenerate primers may possibly amplify regions other than the target due to unspecific binding of the primers to genomic DNA. Multiple infections and unspecific amplification products have been a problem when genotyping by Sanger dideoxy sequencing method and Pyrosequencing technology (paper VI) since a plurality of species/types or unspecific amplification products, present in one specimen, produce sequence signals from all available types/products in the specimen when utilizing a general sequencing primer or one of the amplification primers.

A new approach to address these issues in DNA sequencing technologies is utilizing of target-specific multiple sequencing primers. The new concept offers many advantages, which are explained below.

5.1.5.1. Detection of multiple infections (Paper IV)

Presence of a multitude of species or genotypes in a clinical specimen gives rise to different sequence signals, making typing cumbersome or impossible due to indistinct sequence data. By using target-specific multiple sequencing primers, we have been able to detect and sequence clinically important HPV types. Figure 12 and 13 illustrate the principle of this method. In cases when there are more than two types, a pattern recognition system is used for typing (figure 13).

By pattern recognition is meant comparison-by-alignment of at least two sequence-pattern results and determining the characteristic sequence for each type present in the sample in the sequence pattern combination. Therefore, based on the
Figure 12. When there is a multitude of types/species or/and unspecific amplification products in the amplicon, (a) all the present amplified products will give rise to sequence signals as depicted making typing difficult or impossible. (b) By utilizing target specific multiple sequencing primers, sequencing of unspecific amplifications and irrelevant types can be circumvented resulting in improved and specific sequence signals.
Figure 13. When there is a plurality of types/species or/and unspecific amplification products in the amplified DNA, (a) all the present amplified products will give rise to sequence signals as depicted making typing difficult or impossible. (b) By utilizing target specific multiple sequencing primers, sequencing of unspecific amplifications can be circumvented. In cases when there is a multitude of genotypes, they are sequenced and genotyped by pattern recognition based on sequence characteristic of each type.
number of the oligonucleotide primers applied in the sequencing reaction, the same number of sequence patterns is expected; one characteristic pattern is reserved for each primer used.

Human papillomaviruses (HPV) are the main causative agent of cervical cancer (Munoz and Bosch, 1996; Villa, 1997; Walboomers et al., 1999). It is not an unusual phenomenon for an HPV carrier to be exposed and infected by more than one HPV genotype with the rate of multiple HPV infections varying depending on the characteristics of the population tested. Moreover, the presence of multiple HPV genotypes is a very common incidence in some patient groups. As mentioned earlier, when there is presence of multiple infections in a clinical sample, the general primer used as the sequencing primer will anneal to all genotypes.

We designed seven genotype specific primers with each being specific for a clinical important type. The primers were used in a pool for specific genotyping of the-seven-relevant HPV types. By applying the sequencing primer pool to different samples, we were able to sequence and type these genotypes specifically. This approach allowed us to perform genotyping on samples containing multiple HPV infections by using sequence pattern recognition in pyrograms. The sequencing results were more than satisfactory as we could observe the dominance of the genotypes in multiple infections. In figure 14 we can clearly see the multiple infections of HPV-16 and HPV-18 in three independent clinical samples. HPV-18 is characterized by GGG after the 7th nucleotide addition and HPV-16 is characterized by the sequence peaks for A, C and A, after the 5th, 6th and 9th nucleotide addition, respectively. Figure 14 demonstrates clearly these characteristics in HPV-16 and HPV-18, and simultaneously that the dominance of each genotype can be easily observed in the PCR products. The common/shared and specific bases for each type (noted on top of each peak in figure 14) facilitates genotyping of each DNA target. The dominant type could be easily observed by comparison of single bases shown by arrows. Figure 14a shows HPV-16 and HPV-18 almost in equal dominance, while HPV-18 is dominant in figure 14b and HPV-16 is dominant in figure 14c.
Figure 14. Multiple infections of HPV-16 and HPV-18 in 3 clinical samples sequenced by the multiple sequencing primer method and genotyped utilizing pattern recognition based on characteristic sequence of each genotype. The dark and dotted arrows are intended for demonstrating dominance of HPV-16 and HPV-18, respectively. (a) shows both genotypes almost in equal quantity, (b) HPV-18 is dominant and (c) HPV-16 is dominant.

Pattern recognition can be simplified by a clever choice of primers and selection of proper downstream upcoming sequence for each primer. A pattern differentiation software could be designed to differentiate type-sequence-patterns for relevant human papillomaviruses. In addition, this can also serve as a general principle for microbial and viral typing.

The flexibility in primer design is very advantageous in Pyrosequencing technology. Primers can be designed for specific regions, with the shortest possible sequence-read. This is particularly advantageous by Pyrosequencing technology as the order of nucleotide additions can be easily pre-programmed, and more importantly,
sequencing is performed directly on the first base on the DNA template downstream of the sequencing primer, making primer design more flexible. As mentioned, this eliminates long reads and increases the accuracy. Besides, specific and shorter reads will enhance the sequence quality and detection ability, and it will be time and cost-effective. Many laboratories test only for HPV-16 and HPV-18 genotypes because of their clinical relevance in cervical cancer. As for our assay, HPV-16 and HPV-18 could be easily detected with the first two nucleotide additions of A and C. Figure 15 shows sequence pyrograms of HPV-16 and HPV-18 separately. If the sequence starts with AAC, it distinguishes HPV-16, and sequence ACC characterizes HPV-18, which are sufficient for genotyping and the sequencing time is minimized to a few minutes. Above all, these results are obtained by DNA sequencing, which is far more reliable than hybridization-based techniques.

More importantly, the target specific sequencing primer method could be easily tailored and adapted according to the desired applications or clinical settings based on the regional prevalence of microorganisms and viruses and other associated factors. As the cost for DNA sequencing is decreasing, the same sample could be sequenced in parallel with two or three different target specific primer pools covering a broader spectrum of species/types, or the primer pool of choice could be utilized together with the general primer.

The multiple sequencing primer approach could be used for the Sanger dideoxy sequencing method, although when there is a plurality of relevant types/species in the specimen, pattern recognition is not applicable to electropherograms. However, in such cases, the sequence pattern of the electropherogram reveals multiple infections, which is clinically important. Separate sequencing reactions by each primer are recommended for accurate genotyping. Pattern recognition is suitable for Pyrosequencing technology as it is a non-electrophoretic, bioluminometric method where the sequence signals are detected in real time and there is the possibility of alignment of different sequence signals.
Figure 15. Pyrograms of HPV-16 and HPV-18 sequenced by multiple sequencing primers. The first two nucleotide additions (surrounded by squares) are sufficient for genotyping of these two important high-risk HPV genotypes, minimizing the sequencing time and enhancing the specificity.

5.1.5.2. Unspecific amplification products (Paper IV)

Another limitation for DNA sequencing technologies is presence of unspecific amplifications in the PCR product, making microbial and viral detection a challenging task. General primer sets and other primers aimed for genomic DNA may yield unspecific amplification products (Fernandez-Contreras et al., 2000) making typing difficult in DNA sequencing. This mainly happens when the general amplification primer is used as sequencing primer, which hybridizes to unspecific products present in the sample. The obtained sequence information is complicated or impossible to interpret. Figure 16 shows unspecific amplifications in two ethidium bromide stained agarose gels, in which the products are amplified with general and degenerate primer sets GP5+/6+ and MY09/11, respectively, from genomic DNA. In such cases, nested PCR has been performed, which requires an additional step with no assurance for high-quality results.
Figure 16. Consensus or degenerate primers may yield unspecific amplification products from genomic DNA. Figures (a) and (b) show unspecific amplification on ethidium bromide agarose stained gels. The amplifications were performed by two different general primer sets.

This problem can be solved by using target-specific sequencing primers, which are specifically detecting the species/types of interest as only the primers that have a specific complementary nucleotide sequence in the sample will anneal, and thus, the unspecific amplification products will not result in any sequence signals. Consequently, this approach is highly advantageous as it eliminates the need for nested PCR, cloning or stringent PCR conditions. In our experiments, unspecific amplification products were not observed in amplicons derived from HPV plasmid clones. However, clinical samples showed unspecific amplifications in one-step PCR.
The multiple-sequencing-primer approach improved significantly the sequence data quality for those samples containing unspecific amplification products.

5.1.5.3. Subdominant types and amplicons with low yield

Another limiting factor for DNA sequencing is when the specific sequence signals represent a minority of the total signals from the amplification product and thereby may remain unnoticed. For example, if there are unspecific amplification products in the amplified DNA, the specific sequence signals of the type/species in minority or the specific DNA with low yield will be in the shadow due to dominant and strong signals from the unspecific DNA.

We have been able to address these limitations by utilizing the multiple sequencing primer approach, as the primer annealing is specific for the target DNA and the Pyrosequencing technology has proved to be capable to detect signals from low DNA concentration. This approach eliminates the need for nested PCR or re-performing unsuccessful PCR, sparing time and labor.

The multiple sequencing primer method has shown remarkable results for samples with low PCR yield or when the genotype is in minority compared to sequence signals from unspecific amplification products or multiple infections. Figure 17a shows a challenging clinical sample containing the clinically important high-risk HPV-33, which is in minority (low DNA concentration) and not detectable by the general primer due to high sequence signals from either unspecific amplification products or multiple infections, but is genotyped by multiple sequencing primers (figure 17b). Figure 17c shows the pyrogram of a pure HPV-33 amplicon as a reference for comparison with figure 17b. The challenging template was genotyped correctly despite its low concentration and presence of multiple infections/unspecific amplification products. Sequence pyrogram of each type will reveal easily the genotype (as shown in figure 17b).
Method development and applications of Pyrosequencing technology

**Figure 17.** Detection of a subdominant type with low PCR yield in a challenging clinical sample; sequenced by (a) general primer resulting in unspecific sequence signals and (b) multiple sequencing primer pool of seven primers resulting in specific sequence signals (genotyping HPV-33). (c) Pyrogram of HPV-33 from plasmid for comparison with trace b.

### 5.1.5.4. Faster and more efficient DNA-sequencing (Paper V)

The Pyrosequencing technology, as aforementioned, was earlier limited to sequence information of approximately 20 nucleotides. With the improved chemistry we have been able to achieve DNA read-length of up to 100 bases. The increased read-length has made numerous applications possible, such as microbial typing and
clone checking. Despite the read-length improvements, there are DNA fragments that are challenging to sequence in the expected sequence range due to different factors.

**Figure 18.** Principle of winning read-length by utilizing multiple group specific primers: (a) sequencing by general primer, (b) and (c) sequencing of bacterial group A and B by using group specific sequencing primers annealing downstream of the general primer site in order to win 19 and 31 sequence bases, respectively. The grouping is based on sequence similarities.

We have used the group-specific multiple sequencing primers for winning read-length, increase the sequence quality of challenging templates while decreasing the sequencing time. This strategy is also suitable for samples harboring unspecific amplification products, which may interfere with specific sequencing results.

The principle for winning read-length by using the group-specific multiple sequencing primer approach is demonstrated in figure 18. The primers are specific for a semi-conservative region downstream of the general (amplification) primer, which enables annealing closer to the variable region, allowing further sequencing information from the variable region. The bacterial grouping in these experiments is
based on similarity of sequence alignments of the clinically relevant bacteria studied. In this work, one group of the bacterial strains had sequence similarity in the first 19 bases after the consensus PCR primer and the second group had sequence similarity in the first 31 bases. Based on this information, one group-specific primer was selected for each group, respectively. By using the group-specific primer pool, sequencing of 19 or 31 bases, respectively, could be avoided by annealing downstream of the amplification primer site, which reached closer to variable regions, improved the sequence data quality and decreased sequencing procedure time.

**Figure 19.** Pyrograms from bacterial amplicons sequenced by Pyrosequencing technology utilizing (a) the general primer and (b) the group specific multiple-sequencing primer pool. Trace (b) shows an *E. coli* pyrogram, where 31 semi-conservative bases have been circumvented by using the group-specific multiple-sequencing primers. The arrows indicate positions difficult to read in trace (a) due to unsynchronized extension. This problem has been addressed by using the multiple sequencing primers as shown in trace (b).

Long read sequencing of some templates could be challenging due to reasons such as homopolymeric regions, accumulation of by-products in the system and problematic DNA structures. The above-described approach has proved to be useful for challenging templates that cannot be sequenced in the range expected or templates with unsynchronized extensions. Figure 19a shows a pyrogram obtained from sequencing of *E. coli* using the general sequencing primer. Unsynchronized extension was observed (highlighted by arrows) at the end of the pyrogram. When the group-specific primer pool was utilized (figure 19b), sequencing of a semi-conservative region consisting of 31 bases was avoided. By this approach, information from the variable region was obtained more rapidly. In addition, sequence signals could be
correctly interpreted with no unsynchronized extension. Furthermore, multiple sequencing primers anneal specifically downstream of the general primer site on the DNA template, eliminating the risk for sequencing of non-specific templates. The method was also applicable to the Sanger dideoxy DNA sequencing technique.

5.2. Applications of Pyrosequencing technology (Papers V-IX)

The first applications of the Pyrosequencing method were limited to single nucleotide polymorphisms (SNP) and sequencing short stretches of DNA. New applications have been made feasible due to the advances in the Pyrosequencing chemistry. We have used the enhanced chemistry for microbial and viral typing, mutation detection and EST sequencing.

5.2.1. Microbial and viral typing (Papers V-VII)

Near instantaneous detection of pathogens from clinical material are important for diagnosis, treatment and prophylaxis measurements. The microbial and viral infectious threats have been altered in the course of history as a substantial number of pathogens have been eradicated or controlled. New pathogens are emerging and some are developing resistance. Accurate and specific typing of microbial and viral pathogens is of utmost importance for clinical diagnosis and for characterizing particular types/species/families. Many microorganisms usually lack adequate morphological detail for easy identification. Furthermore, the development of specific therapies/vaccines requires the implementation of sufficient parameters for microbial and viral detection, and a reliable and robust genotyping method is necessary for accurate follow-up during clinical trials and monitoring of treatment (van Doorn et al., 2001).

Pyrosequencing technology, due to its many advantages, has been a useful implement for microbial and viral typing. We have utilized this technique for detection human of papillomaviruses, bacterial and fungal pathogens in clinical specimens.

5.2.1.1. Human papillomavirus genotyping (Paper VI)

Human papillomaviruses (HPV) are part of the Papovaviridae family (Melnick, 1962) with a genome of circular double stranded DNA of approximately 8 kbp. There
are over 100 different HPV types known to date, of which more than 30 infect the cervical mucosa. HPVs are classified on the basis of DNA sequence homology (Chan et al., 1995; Vernon et al., 2000). Definition of a new type is based on sequence homology. A sequence similarity lying under 90% is considered a new HPV type, homology between 90 to 98% constitute a subtype, and a homology above 98% represents a type variant (Morrison and Burk, 1993; de Villiers, 1994).

All the open reading frames (protein coding regions) of HPVs are located on one strand and the genetic organization of the different genotypes is very similar. The HPV genome encodes regulatory and structural proteins called early (E1-E7) and late (L1-L2) proteins. These proteins are involved in viral replication and oncogenicity (Schneider, 1993).

Complete HPV genomic sequences are available for more than half the total number of all identified HPV types. Since type assignment is based on sequence comparisons in the L1 region, the genomic sequences for L1 are present for almost all HPV types on databases. HPVs contain both conservative and highly heterogeneous regions on their genome.

Human papillomaviruses are considered as an important contributing factor in the etiology of certain benign and malignant lesions in humans, such as cervical cancer (Godfroid et al., 1998), and are classified in low- and high-risk genotypes based on their oncogenic potential. High-risk HPVs have been shown to be present in 99.7% of cervical cancer worldwide (Bosch et al., 1995; Walboomers et al., 1999). Therefore, high-risk HPV testing has implications for the clinical management of women with cervical lesions and for primary screening for cervical cancer. In addition, it is necessary to identify individual HPV genotypes to investigate the epidemiology and clinical behavior of particular types (van den Brule et al., 2002). Furthermore, it is not an unusual phenomenon for an HPV carrier to be infected by more than one HPV genotype with the varying rate of multiple HPV infections depending on the characteristics of the population tested. There is preliminary evidence that presence of multiple HPV genotypes may reflect repeated exposure and may also be related to an increased risk for development of disease.

Because all HPV types are closely related, assays can be designed to target conserved regions of the genome or to target regions whose sequences can best be
used to discriminate between different HPV types. Consensus assays use primers directed at relatively conserved regions of the HPV genome. PCR is the most widely used target amplification method and two approaches are most common in HPV testing. One approach utilizes type-specific amplification and the second approach uses a consensus PCR approach. The former approach specifically amplifies a single HPV genotype by the use of type-specific primers. However, the major disadvantage of this approach lies in the fact that many PCR reactions have to be performed. The latter approach utilizes consensus or general PCR primers. There are several different consensus PCR primer sets for amplification of HPV genotypes. The MY09/11 (Hildesheim et al., 1994) and GP5+/6+ (Snijders et al., 1990; de Roda Husman et al., 1995) are two consensus primer sets used widely for detection of a broad-spectrum of HPV genotypes. These primers amplify fragments on the L1 region of HPV yielding amplicons of 450 and 150 bp, respectively.

We have sequenced clinical HPV samples by Pyrosequencing technology to determine the HPV type. We used the two commonly used primer systems, the MY09/11 primers and the GP5+/6+ primers, which amplify a broad spectrum of HPV genotypes (Resnick et al., 1990; de Roda Husman et al., 1995). GP5+ general primer was applied as sequencing primer and the first 14 to 25 bases downstream of this primer were adequate for accurate genotyping. The obtained sequence information was analyzed in the GenBank database for genotyping. For samples containing unspecific amplification products nested PCR was performed. The problem with multiple infections and unspecific amplification products were dealt with by using the multiple sequencing primer approach as described earlier. Pyrosequencing technology proved to be a very reliable and rapid typing method with the highest discriminatory ability for HPV detection.

5.2.1.2. Bacterial identification (Paper V)

Nucleic acid based analytical and diagnostic methods have revolutionized molecular differentiation of microorganisms. Phenotypic typing approaches are gradually being supplemented or substituted by new genotyping methods allowing a more distinguished and detailed discrimination of closely related microorganisms, pathogens, bacterial strains and isolates on the DNA level.
The 16S rRNA is a structural part of the ribosomal subunit, which has been generally used for analysis of phylogenetic relationships between bacteria. The 16S rDNA is approximately 1500 bp long, consisting of eight highly conserved regions (U1-U8) and nine variable regions (V1-V9) in the bacterial domain (Gray et al., 1984). These features are not only valuable for bacterial phylogenetic studies but also for bacterial detection and identification in the clinical laboratory (Van de Peer et al., 1996).

Several features of the 16S rRNA gene have made it a useful target for clinical identification (Patel, 2001). Firstly, it is present in all bacteria making it a universal target for bacterial identification. Secondly, the function of 16S rRNA has remained constant over a long period, so sequence changes are more likely to reflect random changes than selected changes, which would alter the molecule’s function (Woese, 1987). And moreover, the 16S rRNA gene is large enough to contain statistically relevant sequence information.

DNA sequencing of the entire 16S gene is not practical for clinical diagnostics as the present sequencing technologies can offer read-lengths of 500 to 650 bases, meaning that several sequencing reactions must be performed to sequence the entire gene, which is labor-intensive and expensive. Fortunately, for most bacteria, the entire gene does not have to be sequenced to achieve the species-level. Even though phylogenetical informative positions are lying throughout the entire 16 rRNA gene, the region with the greatest sequence heterogeneity lies approximately in the first 500 sequence bases of the 5’-end (Woese et al., 1983; Rogall et al., 1990; Tang et al., 1998; Harris and Hartley, 2003). Furthermore, extensive databases of rDNA sequences are becoming available for all types of microorganisms (Benson et al., 2000), allowing the design of oligonucleotides that can differentiate clinically important species from all known microorganisms.

For bacterial detection and identification, universal/consensus primers that bind to conservative regions, are often used for amplification (Woese, 1987; Larsen et al., 1993; Maidak et al., 1994). Since the 16S rRNA genes of almost all common bacterial pathogens found in body fluids have been sequenced, it is possible to design PCR primers capable of amplifying all eubacteria based on the conservative nature of the 16S rRNA genes (Greisen et al., 1994; Radstrom et al., 1994; McCabe et al., 1995).
Furthermore, the 16S ribosomal RNA-coding sequences contains regions that are found in all bacteria but not in human DNA, and these regions flank species-specific variable sequences (McCabe et al., 1995).

For bacterial detection, we performed partial 16S rDNA amplification by utilizing a universal primer pair, followed by DNA sequencing in a blind test by Pyrosequencing technology for family/species typing of a selected bacterial test panel relevant for neonatal severe infections. We chose the V6 region, as it is one of the most highly variable regions on the 16S rDNA. The objective was to obtain sequence data of up to 60 bases for bacterial identification as 36 to 55 bases were needed for family identification by using the general sequence primer. Moreover, three species were typed to species level with sequence data of 50 bases (Haemophilus influenzae, Neisseria meningitidis and Streptococcus agalactiae) and one with 68 bases (Staphylococcus haemolyticus).

This was the first study performed with the improved Pyrosequencing technology for sequencing of extended read-length of a clinical test panel. Sequence data in the range 50 to 110 bases were obtained depending on the bacterial strain (a few challenging templates could not be sequenced over 50 bases). As explained earlier, the multiple sequencing primer approach could circumvent 19 and 31 bases dependent on the bacteria sequenced, allowing to genotype more easily challenging bacterial species. The obtained results were satisfactory. Figure 20 demonstrates pyrograms of 3 bacterial strains sequenced by general primer.

The bacterial identification by Pyrosequencing can be developed further and applied to other regions for different bacterial targets and can be a useful utility in diagnostics and clinical laboratories.
5.2.1.3. Fungal identification (Paper VII)

Human pathogenic fungi and yeasts are ubiquitous in the environment, of which some of them are normal inhabitants in the body. They are usually opportunistic organisms, causing acute-to-chronic infections when conditions in the host are favorable (Mannarelli and Kurtzman, 1998). The incidence of life-threatening systemic fungal infections has been increasing in recent years, and the increasing incidence has been correlated with increasing numbers of immunocompromised patients (Anaissie et al., 1989).

Drug resistance is a major problem in treating yeast infections (White et al., 1998). Therefore, quick and accurate identification of disease-causing yeasts is important as it allows the delivery of the most effective drug and the use of the proper dose of drug for the particular infection. In addition, a molecular detection approach that can give definitive identification with same-day results and provide valuable information to physicians for patient management is clinically favorable.

Like bacteria, rRNA genes are good candidates for fungal detection. 18S rRNA genes are attractive targets for fungal and yeast amplification-based detection assays,
since these genes are present at a high copy number, thus increasing the sensitivity of the PCR (Hopfer et al., 1993). Furthermore, rRNA genes are composed of regions of higher and lower evolutionary conservation, thereby enabling amplification at different taxonomic levels. Consequently, nuclear ribosomal genes have been widely used as a target for detection of fungal microorganisms by molecular methods (Hopfer et al., 1993; Makimura et al., 1994; Haynes et al., 1995).

The rates of isolation of the major yeast species causing fungemia have been determined in several studies: Candida albicans (50 to 59%), Candida tropicalis (11 to 25%), Candida glabrata (8 to 18%), Candida parapsilosis (7 to 15%), Candida krusei (2 to 4%), Candida neoformans (~2%), and other species (~2%) (Beck-Sague and Jarvis, 1993; Rex et al., 1995a; Rex et al., 1995b; Nguyen et al., 1996; Pfaller et al., 1999). Therefore, earlier information regarding the species causing fungemia is valuable to help physicians to select appropriate antifungal agents and regimens to treat patients. The rate of isolation of C. glabrata from blood cultures has increased from 8% to about 20% during the period from 1952 to 1992 according to recent surveys (Pfaller et al., 1998; Pfaller et al., 1999). This increase might be due to the widespread use of antifungal drugs for prophylaxis and treatment of candidiasis, affirming the need for more rapid and accurate identification.

We utilized the Pyrosequencing technology for typing of fungal isolates and clinical samples from immunocompromised patients suffering from proven invasive fungal infections. The DNA was amplified by general consensus primers (Einsele et al., 1997), which were based on a highly conserved region within the 18S rRNA gene allowing the amplification of a broad range of fungal pathogens. The amplified DNA fragments were sequenced up to 40 bases to identify the samples. All the samples were accurately identified by the Pyrosequencing technology. The reproducibility of the technique was confirmed by sequencing the amplicons several times with identical sequence data each time for every sample. The sequence data obtained by the Pyrosequencing technology suggested that 18 to 32 bases were sufficient for identification of C. albicans, C. glabrata, C. krusei, C. parapsilosis, C. tropicalis and Aspergillus subsp. The species that we identified in our assay covers a wide range of fungal infections as fungemia is usually caused by a limited number of fungal species. In fact, the most commonly encountered yeast species (C. albicans, C. tropicalis,
C. glabrata, C. parapsilosis, C. krusei, and C. neoformans) may represent 95% of the total yeasts recovered from blood cultures (Rex et al., 1994; Nguyen et al., 1996; Pfaller et al., 1999; Chang et al., 2001). According to our results, Pyrosequencing technology appears to be a good diagnostic tool for detection and identification of fungal pathogens. Fungal identification can be developed further for Pyrosequencing technology.

5.2.2. EST sequencing (Paper VIII)

The sequence order of nucleotides determines the nature of the DNA. A relatively straightforward approach to identify and characterize genes is to isolate mRNA from different tissues at different ages, convert the mRNA to complementary DNA (cDNA) and sequence single reads of cDNAs, also called Expressed Sequence Tags or ESTs. This approach leads to the discovery of genes and their relative expression levels (Okubo et al., 1992; Adams et al., 1993; Adams et al., 1995; Braren et al., 1997).

Theoretically, eight or nine nucleotides in a row should define a unique sequence for every gene in the human genome. However, it has been found that in order to uniquely identify a gene from a complex pool, a longer sequence of DNA is required.

In order to perform large-scale EST analysis robust, accurate, and inexpensive platforms allowing time-effective analysis of large numbers of samples will be required.

To increase the capacity of conventional DNA sequencing in comparative EST analysis, serial gene expression analysis (SAGE) (Velculescu et al., 1995; Brenner et al., 2000) was introduced. The key principle of SAGE lies in the fact that a very short sequence tag of a transcript is required to uniquely identify the origin of the transcript on EST databases. These short cDNA tags are 9-13 nucleotides long. Compared to single-pass partial sequencing methods, SAGE produces approximately 25-30 times more gene-specific sequence data of several gene-tags in a single run. However, this technique faces limitations, since many mRNA molecules do not have proper restriction sites for cleavage by the conventional endonucleases. Furthermore, the accuracy of gene identification from the obtained sequence length (9-13 bases) is limited to about 90-95 percent. In a study utilizing SAGE, the accuracy of gene identification in yeast was about 90 percent (Velculescu et al., 1997).
In our pilot study, it was found that 98% of all genes could be uniquely identified by sequencing a length of 30 nucleotides. Pyrosequencing technology was used to sequence for EST sequencing and consequently, gene identification. The results were in complete agreement with sequence information that was obtained by the dideoxy method.

The present improved Pyrosequencing chemistry and automation allows increased read-length sequencing and screening of cDNA libraries in a high-throughput fashion.

5.2.3. SNP analysis (Paper IX)

Single nucleotide polymorphisms (SNPs) are a molecular form of genetic variation (Collins et al., 1997). Because of their ubiquity, there are high frequency SNPs in all human populations and they are plentiful in the genome. Present estimates are that SNPs with minor allele frequencies of at least 1% are found at the rate of 1 in every 200–300 bp in the genome (Kruglyak and Nickerson, 2001; Stephens et al., 2001). Extrapolating to the entire genome, this suggests that perhaps as many as 15 million SNPs are ultimately available to study (Salisbury et al., 2003). For correlation of specific phenotypes to genomic variation, such as disease susceptibility or variable drug response, it is important to have knowledge of how the variation is organized and distributed among genes, gene regions, and populations. Because of their dense distribution across the genome, SNPs are viewed as ideal markers for large-scale genome-wide association studies to discover genes in common complex diseases. When SNPs are discovered, they are assembled into haplotypes (the set of polymorphism alleles that co-occur on a chromosome) to examine allelic variation.

As the number of identified SNPs escalates, there will be an increasing demand for robust and accurate methods to type and assess the biological impacts of these genetic variations. The golden standard method for SNP scoring is DNA sequencing. In this work, we have applied the Pyrosequencing technology for SNP detection for the very first time (also Nordstrom et al., 2000b for double-stranded SNP analysis). The Pyrosequencing method was used for sequencing of four SNPs located on chromosomes 9q and 17p from 24 unrelated individuals, resulting in accurate results. A unique property of Pyrosequencing technology in SNP typing is that each allele combination (homozygous, heterozygous) will give a specific pattern compared to the
two other variants. This important feature makes typing extremely accurate and easy.

SNP typing and analysis is one of the major applications of Pyrosequencing technology suitable for automation and high-throughput analysis.
6. Future trends of Pyrosequencing technology

DNA sequencing is one of the most central platforms for the study of biological systems. Future applications require more robust and efficient DNA sequencing techniques for sequence determination. As described in this thesis, Pyrosequencing technology is a relatively novel DNA sequencing method still at its infancy. Being a young technique, it holds numerous unique advantages, which makes this method a powerful tool in the field of DNA sequencing technologies.

Conventional DNA sequencing, first described over 26 years ago (Sanger et al., 1977), has gone through major improvements by a large number of academic and research groups as well as commercial companies during these years. The Pyrosequencing method has not been exposed to such concentration of research yet. There is extensive space for developments and improvements in chemistry and instrumentation.

The Pyrosequencing method has already shown evidence of high accuracy in DNA sequencing and analysis of polymorphic DNA fragments in many clinical and research settings. The technique in many ways is already competitive in comparison to other existing sequencing methods. By increasing the read length to higher scores and by shortening the sequence reaction time per base calling, Pyrosequencing can take a more broad position in the DNA sequencing arena. Miniaturization of this technique, and its integration with nucleic acid amplification and fully automated template preparation is envisioned for future application formats, as the trend is directed to analysis of lesser amounts of specimens and large-scale settings, with higher throughput and lower cost. Microfluidic formats appear to fulfill these requirements, as most importantly, the problem with product accumulation will be minimized due to constant removal of unincorporated nucleotides, resulting in more synchronized extension.
7. Concluding remarks

The Pyrosequencing method has emerged as a versatile DNA sequencing technology suitable for numerous applications. It is the first alternative to the conventional Sanger dideoxy method for de novo DNA sequencing. This technique was earlier restricted to sequencing and analysis of SNPs and short stretches of DNA. However, the Pyrosequencing technology has gone through phenomenal developments and improvements both in chemistry and instrumentation. Indeed, there is room for further advancements as only a couple of academic research groups have focused on this method.

By removal of inhibitory factors and improved chemistry we have been able to enhance the sequence quality and read-length, which has opened new avenues for many new applications in de novo sequencing, as well as, in re-sequencing, mutation detection, and microbial and viral typing. The novel multiple sequencing primer method has also contributed to typing of samples containing a multitude of types/species and unspecific amplification products in clinical settings, eliminating the need for stringent PCR reactions, nested PCRs and cloning. We believe that the multiple sequencing primer strategy has many other potential applications especially as automation is being more and more integrated into clinical settings and the cost for DNA sequencing is dropping.

Further advancements in Pyrosequencing technology will be to achieve longer read length for applications requiring DNA reads over the present boundary, to reduce the sequencing time frame and to decrease the sample quantity and further improvements in automation.

Completion of the human genome does not imply that no further sequencing efforts are necessary. Without doubt, the next technological development is directed to generate fast-sequencing in microfabricated formats in order to bring the power and accuracy of sequencing analysis in diagnostics to hospitals and clinical laboratories.
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADP</td>
<td>adenosine diphosphate</td>
</tr>
<tr>
<td>AMP</td>
<td>adenosine monophosphate</td>
</tr>
<tr>
<td>APS</td>
<td>adenosine phosphosulfate</td>
</tr>
<tr>
<td>ATP</td>
<td>adenosine triphosphate</td>
</tr>
<tr>
<td>bp</td>
<td>base pair</td>
</tr>
<tr>
<td>CCD</td>
<td>charge-coupled device</td>
</tr>
<tr>
<td>cDNA</td>
<td>complementary DNA</td>
</tr>
<tr>
<td>CE</td>
<td>capillary electrophoresis</td>
</tr>
<tr>
<td>dATP</td>
<td>deoxyadenosine triphosphate</td>
</tr>
<tr>
<td>dATP[S]</td>
<td>deoxyadenosine alfa-thio triphosphate</td>
</tr>
<tr>
<td>ddATP</td>
<td>dideoxyadenosine triphosphate</td>
</tr>
<tr>
<td>ddCTP</td>
<td>dideoxycytidine triphosphate</td>
</tr>
<tr>
<td>ddGTP</td>
<td>dideoxyguanine triphosphate</td>
</tr>
<tr>
<td>ddNTP</td>
<td>dideoxynucleotide triphosphate</td>
</tr>
<tr>
<td>ddTTP</td>
<td>dideoxythymidine triphosphate</td>
</tr>
<tr>
<td>dGTP</td>
<td>deoxyguanine triphosphate</td>
</tr>
<tr>
<td>dNDP</td>
<td>deoxynucleotide diphosphate</td>
</tr>
<tr>
<td>dNMP</td>
<td>deoxynucleotide monophosphate</td>
</tr>
<tr>
<td>dNTP</td>
<td>deoxynucleotide triphosphate</td>
</tr>
<tr>
<td>ds</td>
<td>double-stranded</td>
</tr>
<tr>
<td>dTTP</td>
<td>deoxythymidine triphosphate</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>EDTA</td>
<td>ethylenediamine tetraacetic acid</td>
</tr>
<tr>
<td>EST</td>
<td>expressed sequence tag</td>
</tr>
<tr>
<td>HPV</td>
<td>Human papillomavirus</td>
</tr>
<tr>
<td>kb</td>
<td>kilobase</td>
</tr>
<tr>
<td>KF</td>
<td>Klenow fragment</td>
</tr>
<tr>
<td>$K_m$</td>
<td>Michaelis-Menten constant</td>
</tr>
<tr>
<td>$k_{cat}$</td>
<td>catalytic constant</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>MS</td>
<td>mass spectrometry</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>PPi</td>
<td>inorganic pyrophosphate</td>
</tr>
<tr>
<td>SBH</td>
<td>Sequencing-by-hybridization</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------------------------</td>
</tr>
<tr>
<td>SBS</td>
<td>Sequencing-by-synthesis</td>
</tr>
<tr>
<td>SNP</td>
<td>single nucleotide polymorphism</td>
</tr>
<tr>
<td>ss</td>
<td>single-stranded</td>
</tr>
<tr>
<td>SSB</td>
<td>single-stranded DNA-binding protein</td>
</tr>
<tr>
<td>TA</td>
<td>Tris-acetate</td>
</tr>
</tbody>
</table>
Acknowledgements

This thesis is a synergistic result of many minds, collaborations and fruitful meetings and dialogues. I am deeply grateful for the inspiration and scientific sources of many scholars and thinkers and for the trans-generational knowledge and wisdom. I take this opportunity to express my profound gratitude to the entire staff at the biotechnology department for productive collaborations. I would like to thank in particular:

Pål Nyrén, for his admirable supervision and indulging me the freedom and support to work on my areas of interest. I would like to thank him for his scientific guidance, for his sense for constructive details, having always time for me, efficiently dealing with matters, creating a pleasant working atmosphere, and simply, for being an altruistic fine scientist.

Håkan Mellstedt, my very first supervisor and a “sui generis”, whom I have always looked up to as a remarkable scholar and scientist, an industrious group leader blessed with virtues of determination, modesty, patience and warm heart. I am also very grateful for all the helpful insights and for always gratifying me his precious time despite his tremendously busy schedule.

Mostafa Ronaghi, one of the senior members of our group, for great and supportive friendship, for initiating my collaboration with Pål and Pyrosequencing method, for his ceaseless enthusiasm, hyperactive vision and most of all for his general generosity and well-acknowledged positive attitude.

Nader Nourizad, a great friend, colleague and our protein expert who has made our lab a fun place with his great sense of humor and happy-go-lucky attitude. Thank you for assistance, collaboration and all the interesting conversations. Tommy Nordström, for good collaboration and most of all for your witty sense of humor and your modesty. Jonas Eriksson, for all the nice conversations and collaborations. Carlos Garcia and Samer Kara, the old members of the group, for your friendship and collaboration.

Pål Nyrén, Liam Good and Mostafa Ronaghi for their valuable comments on this thesis.

Afshin Ahmadian, for friendship and for all the enlightening scientific conversations, always taking your time with patience to have fruitful dialogues, which have assisted me in my work.

Bahram Amini and Mehran Gharderi, for your friendship, fruitful collaborations and for always being ready to lend a hand.

Nader Pourmand, my best friend and my college classmate. Thank you for all your support and help, your warm hospitality and generosity during all these years.

Alphonsa Lourdudoss, Anne Jifält, Lotta Rosenfeldt, Marita Johnsson, Monica Ruck and Pia Ahnlén for all excellent professional assistance during the years, and Tommy Östlund, our nirious technical person, for his well-needed technical assistance.

Anders Holmberg and Morten Lukacs for their friendliness and excellent assistance with the magnetic robotic technique.
All the new and old members of Håkan mellstedt's group, particularly the very senior members and the heart of the group Birgitta Hagström, Gun Ersmark, Ingrid Eriksson and Lena Virving who welcomed me and taught me the basic laboratory disciplines when I started in IGT lab in 1993. You have always been friendly, supportive and helpful. I would like to thank Barbro Näsman-Glaser and Anna-Lena Borg who later joined the group.

Gunilla Burén, for great professional assistance, amazing excellent efficiency and a warm heart. Gerd Ståhlberg, for efficient professional and technical assistance and friendly attitude. Anna Lena Hjelm-Skoog and Peter Ragnhammar, for good collaboration, sympathy and for your medical expertise. Anders Eklöf, Joe Lawrence, Sören Lindén. Max Pettersson at CCK and Ricardo Giscombe at CMM for admirable technical and "helping-out" assistance and for your friendliness.

The staff and researchers at Cancer Center Karolinska (CMM) and Center for Molecular Medicine (CMM).

Keng-Ling Wallin and Biying Zheng, Joakim Dillner, Lena Klingspor and her group, Nongnit Lewin and her group, Ann Kari Lefvert and Xiao Yan Zhao, Bo Johansson and Mina Kalantari, Per Olcén and his group, Bengt Wretlind and Emma Lindbäck, Maria Oggionni, and Edit Akom for fruitful collaborations.

The staff at Pyrosequencing AB for technical assistance and expertise, particularly Björn Ekström, Björn Ingemarsson, Anders Alderborn, Thomas Hultman, Nigel Tooke, Peter Hagerlid, Peter Lindqvist, Christina Johansson, Jenny Dunker, Kristin Hellman, Annika Tallsjö and Per-Johan Ulfendahl.

Cecilia Beer, my omniscient friend, for all the illuminating cultural, lingual, art and history conversations. Carin beer for your hospitality.


Bijan Tousi, my best friend in Toronto. You have been a true devoted and loyal comrade through the years. I am not sure if I can ever thank you enough.

Kiriaki Ifantidou for being an incredible mother and always prioritizing Jonathan con amore. George, Eleni, Costas, Evi, Gregorios, Maria, Nicoleta and Margareta for friendship, hospitality and for all the wineing and dining during my stays in Greece.

My parents, my sister Mina and my brother Ciamack for their constant love and support.

Jonathan, my little prince, my teacher and my source of joy and happiness.
References


Method development and applications of Pyrosequencing technology


