Reductionist approach to avoid information overflow and noise in the assessment of complex DNA mixtures

Dissertation
zur Erlangung des Grades
des Doktors der Naturwissenschaften
der Naturwissenschaftlich-Technischen Fakultät II
- Physik und Mechatronik -
der Universität des Saarlandes

Von

Bokkasam Harish

Saarbrücken

(2013)

Tag des kolloquiums: 9.12.13

Dekan: Univ.-Prof. Dr. Christian Wagner

Mitglieder des Prüfungsausschusses:

Vorsitzende:	Univ.-Prof. Dr. Ralf Seemann
Berichterstatter:	Univ.-Prof. Dr. Albrecht Ott
	Prof. Dr. Ingolf Bernhardt

Akadamische mitarbeiter: Dr. Haibin Gao

Zusammenfassung

Genetische Informationsübertragung ist die Übertragung der elterlichen Eigenschaften auf die Nachkommen, sie hängt vollständig von spezifischer molekularer Erkennung ab. Die Information ist in bestimmten DNA Abschnitten (Gene) gespeichert. Verschiedene Genexpressionsanalyse-Studien werden zur Identifizierung und Quantifizierung der Aktivität bestimmter Gene durchgeführt. Viele Studien konnten erfolgreich zeigen, dass die "molekulare Erkennung“ von der Komplexität der biologischen Systeme abhängig ist. Eine Erhöhung der Komplexität führt häufig zu unspezifischer Hybridisierung. Die anschließende Quantifizierung dieser Daten führt zu Ergebnissen verringertem Signal-Rausch-Verhältnis.

Oft sind die angestrebten Informationen in größeren Mengen an unspezifischen Informationen verborgen, insbesondere wenn es sich bei den gesuchten Informationen um kleine Datenmengen handelt. Oft stellen aber große Mengen von unspezifischen Informationen eine bedeutende Herausforderung dar, indem sie die Aufwand zur Informationsgewinnung erheblich erhöhen.

Der erster Teil der vorliegenden Dissertation beschäftigt sich mit der Entwicklung eines molekularbiologischen Modellsystem. Der zweite Teil befasst sich mit der Identifizierung, dem Abruf und der Überprüfung der spezifischen Informationen aus dem Modellsystem durch Hybridisierung. Schließlich werden die gewonnenen Informationen mittels verschiedener Genexpressionanalysetechniken auf ihre Genauigkeit und Spezifität hin untersucht.

Abstract

Genetic information transfer, which is the transfer of parental traits to offspring’s, completely depends on specific molecular recognition. This information is stored in stretches of DNA known as genes. Various gene expression profiling studies are used to investigate the activity of specific genes at specific instant. Many studies have successfully shown that molecular recognition is dependent on the complexity of biological systems. Increase in complexity often results in unspecific hybridization. Subsequent quantification of this data results in reduced signal to noise ratio.

Often, the information of interest, especially when it is in small amounts, is clouded by larger amounts of unspecific information. These problems need to be addressed for profiling studies to be more reliable. Often large amounts of unspecific information may present a significant challenge, as it increases the time and effort needed to retrieve specific information.

This dissertation proposes a novel and simple method to identify and reduce hybridization noise in gene expression profiling techniques. First part of this dissertation deals with the development of a biomolecular model system. Second part deals with identification, retrieval and subsequent verification of the specific information from model system with noise. Finally, the retrieved information is analysed for fidelity and specificity by various gene expression profiling techniques.

Acknowledgements

It is only apt to begin this dissertation by acknowledging the support of many people, I was fortunate to be acquainted with during my doctoral work.

First and foremost person to whom I am always thankful for making my doctoral dissertation possible will be Prof. Dr. Albrecht Ott. I was coming from a different academic background and time and support he provided me during the initial stages in Bayreuth and in Saarbruecken deserves a special mention. Throughout my doctoral dissertation period, he was always available for discussions and necessary guidance. Also, sudden and fruitful ideas during coffee and lunch breaks were quite helpful.

My colleagues in Bayreuth and Saarbruecken made my time during my dissertation a memorable experience. Everyone in Bayreuth was quite helpful during initial stages. In Saarbruecken, I was again fortunate to have wonderful colleagues. Heike Wech was always quite patient with my questions in biology. I am also happy to count Eva Wollrab and Manuel Worst among my friends. The detailed and extensive proofreading from Dr. Mikhail Zhukovsky (Mischa) made my dissertation much better. Christian Trapp was quite supportive along with Marc Schenkelberger.

Dr. Sascha Tierling and Dr. Konstantin Lepikov were quite helpful and resourceful with my needs and queries in experimental part.

My Indian friends in general and Telugu people in particular made my stay in Saarbruecken a nice experience to remember. Especially, Shivasanker Lingam was quiet patient with my questions in biology and genetics. I am always thankful for his support. Aravind Pasula and

Srikanth Duddela made my occasional coffee breaks into a necessity with their fun and wit. Shiva and Srikanth made Friday nights get together into a wonderful experience.

My family have supported me throughout my life. Especially my parents (B. Ramachandra Rao & B. Ramadevi) and relatives (P. Rukmini & P. Ramprabhu) are responsible for making my higher studies in Germany possible with financial and moral support.

It is proper to mention one person who deserves much credit for successful completion of my doctoral dissertation. I will always be proud to have Prathyusha (bujji) as my wife. She has supported me during most important period of my life. Her presence and not to forget excellent food made my writing part of the dissertation into a wonderful experience.

There are many more people who have supported me in Germany. I thank all of them for making my stay in Germany into a memorable experience.

Table of Contents

1. Introduction

Molecular recognition plays a very important role in evolution of biological systems. It is the basis for all processes at molecular, cellular and organizational levels. At molecular level, this process is linked to interaction between host molecule and guest molecule to form a host-guest complex. Genetic information transfer, which is the transfer of parental traits to offspring's, completely depends on specific molecular recognition. This information is stored in stretches of DNA known as genes. Various gene expression profiling studies are used to identify and quantify the activity of specific genes at specific instant.

In gene expression profiling studies, various techniques are used to identify and quantify information present in specific regions of DNA. These are generally categorised into random fragmentation methods which break a biological template at random points or Polymerase Chain Reaction (PCR)-isolation of specific regions of biological template. The obtained information is analysed by various characterization techniques. Many of these techniques use various biomolecular recognition processes like ligand-receptor complex formation, protein-antibody binding, nucleic acid hybridization etc. Recent advances in genetics and evolution of bioinformatics as data processing systems has increased the importance of high throughput applications like DNA microarrays and quantitative PCR. These techniques use base pairing which is hybridization/binding of complementary nucleic acid sequences.

Many studies have successfully shown that molecular recognition is dependent on the complexity of biological systems. Increase in complexity often results in unspecific hybridization. Subsequent quantification of this data results in hybridization noise and reduced signal to noise ratio. Understanding and eliminating unspecific hybridization is required for all successful gene expression profiling studies.

Previous research has sought to find solutions to problems involving discrepancies in gene expression profiling studies. Most of the experimental studies used brute force techniques like random fragmentation and random sequence amplification methods. These methods produce large amounts of data that has to be evaluated/analysed/sequenced for information gathering. Often, the information of interest, especially when it is in small amounts/data sets, is clouded by larger amounts of unspecific information. Once the specific information is irreversibly lost, it cannot be balanced using advanced statistical models. These problems need to be addressed for profiling studies to be more reliable. Existing bioinformatics capabilities provide ways to analyse large amounts of data produced by existing sample preparation techniques.

Often large amounts of unspecific information may present a significant challenge for even latest and improved bioinformatics tools, as it increases the time and effort needed to retrieve specific information. This leads to problems associated with existing data analysis tools. So there is a need to produce reduced amounts of specific data which in turn reduces the need and problems with time consuming data analysis to retrieve specific information.

This dissertation proposes a novel and simple method to identify and reduce hybridization noise in gene expression profiling techniques. First part of this dissertation deals with the development of a biomolecular model system, which has specific information along with unspecific information of high complexity. This adds noise sources to our model system. Second part deals with identification, retrieval and subsequent verification of the specific information from model system with noise. Various biochemical and molecular genetics techniques are used for this purpose. Finally, the retrieved information is analysed for fidelity and specificity by various gene expression profiling techniques.

1.1. Background and fundamentals

1.1.1. Central Dogma of Molecular Biology

Marshall Nirenberg *said, "DNA makes RNA makes protein.*

The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to either protein or nucleic acid (Crick, 1970)

This concept mainly explains sequential transfer of information between bio-polymers. In general, these bio-polymers are biological specimens. DNA, RNA (both nucleic acids) and proteins are three important classes in bio-polymers Nucleic acids are built by nucleotides.

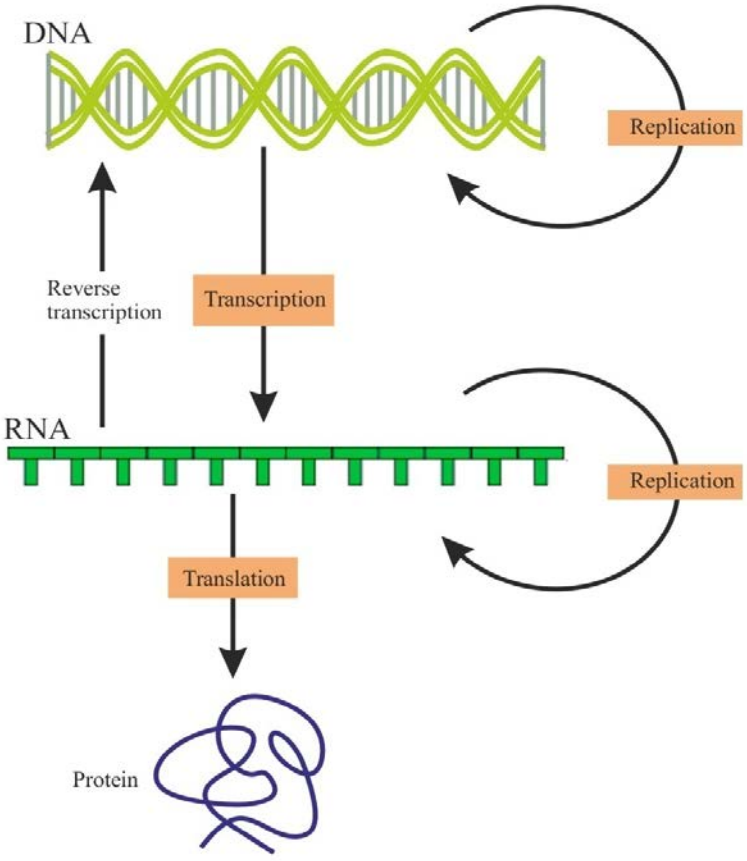


Figure 1.1: Central Dogma of Molecular Biology

Figure 1 shows a depiction of information flow according to central dogma of molecular biology. According to central dogma, sequential information of an individual which is present in the DNA of that individual is transcribed into RNA through a process known as transcription. Then the information present in RNA is finally translated into an active protein. These transfers can be placed into three classes. These three classes of sequential information transfer were proposed by Crick. These transfers are tabulated in the Table 1.1.

Table of the 3 classes of information transfer suggested by the dogma

	General		**Special**		**Unknown**
	DNA → DNA		RNA → DNA		protein → DNA
	DNA → RNA		RNA → RNA		protein → RNA
	RNA → Protein		DNA → Protein		Protein → Protein

Table 1.1: Classes of sequential information transfer (Crick, 1970)

- **General transfer**

  Common and conventional transfer of sequential information can be classified as general transfers. These general transfers include DNA replication, transcription and translation.

  1. DNA replication - DNA copied into identical DNA sequence.

  2. Transcription -Sequential information in DNA is copied into an intermediate form which is messenger RNA (mRNA). RNA polymerase and transcription factors are involved in transcription.

  3. Translation- Finally active proteins can be synthesized using the sequential information present in mRNA. The mRNA sequence functions as a template for protein synthesis (Crick, 1970).

- **Special transfer**

  Reverse transcription is the transfer of sequential information present in RNA to DNA. This process occurs in retroviruses.

- **RNA replication**

  In RNA replication, specific RNA sequence is copied into another RNA sequence. This is known to happen in viruses (Hacker, 1992).

- **Direct translation from DNA to protein**

  Direct translation from DNA to protein has been demonstrated in a cell-free system. This is performed using *E. Coli* extracts that contact all the ingredients for protein synthesis (in the absence of cells) (Kobayashi et al., 2007).

DNA, RNA and protein can be classified as linear polymers. According to central dogma, the sequence of one bio-polymer, either DNA or RNA is used to develop a sequence of another bio-polymer. Deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) are the basic nucleic acids (Pieztch, 2003). Figure 1.2 shows the DNA sequence in coiled structure.

Nucleotides and **nucleic acids** are biological molecules which contain hetero cyclic nitrogenous bases that form the core components. The biochemical roles of nucleotides are numerous; they participate as essential intermediates in virtually all aspects of cellular metabolism. Linear polymers of amino acids form proteins; linear polymers of nucleotides form nucleic acids.

This means the original sequence serves as a template that stores information (Goldman et al., 2013). The whole transfer mechanisms are completely governed by the complexity of the original template.

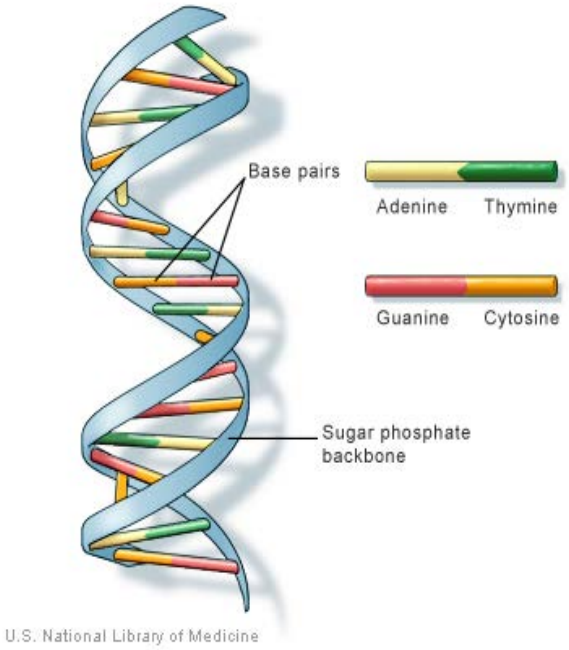


Figure 1.2: DNA structure

The three general transfers, replication, transcription and translation form the backbone of sequential information flow in any biological system. Depending on the complexity of the biological system, these general transfers can also occur simultaneously. Also, information in a single gene can be used to synthesize different proteins. The information flow is also influenced by various external and internal factors. Some proteins are synthesized only when some particular stress factors are induced. For example, exposure of human body to UV light increases the production of Melanin pigment. This protects the skin from possible damage.

1.1.2. Gene

A gene is can be described as a stretch of DNA or RNA. These stretches have the code for a RNA chain (in case of DNA) and for a polypeptide chain (in case of RNA). These codes determine the function of an organism. Gene are very important in biological systems

(Missiuro et al., 2009). They specify the role of all the proteins and functional RNA. The information to construct and sustain an organism is present in the genes. This information transferred as hereditary trait form parent to offspring. One single gene can form the basis for synthesis multiple proteins for different functions. The nucleotide composition of a gene is shown in Figure 1.3.

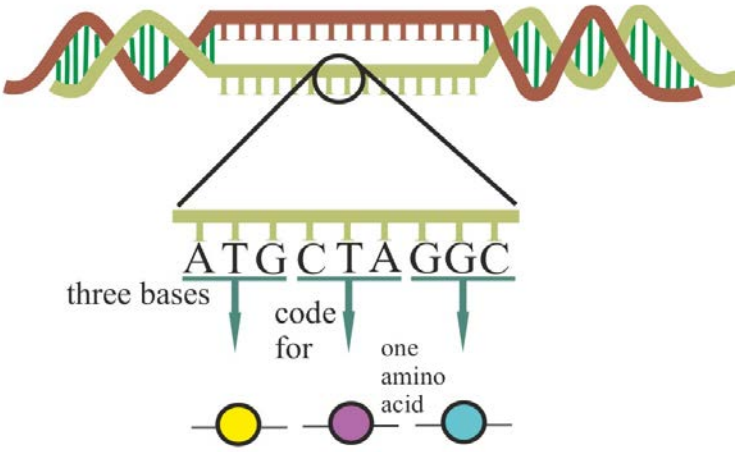


Figure 1.3: Coding information in a Gene

1.1.3. Molecular recognition

Molecular recognition is very important in the evolution of biological systems. Molecular recognition plays in an important in all the processes that are mentioned in central dogma of molecular biology. Due to this process, various biomolecules recognize their subjective partner molecule (Graham et al.,, 2003). This results in subsequent transfer of genetic information. Thus molecular recognition is central to all processes involving information transfer. Specific molecular recognition is necessary for accurate transfer of genetic information or hereditary traits from individual to offspring. Understanding molecular recognition is important to analyze and investigate hereditary diseases. Gene expression profiling is based on specific molecular recognition (Tarca et al., 2006). All the gene expression profiling techniques are developed from basis on some form of molecular

recognition.

Specific molecular recognition depends on DNA catalysis. Due to the catalytic action two complementary strands of DNA present in a solution, hybridize (recognize each other) to form a DNA-DNA duplex. Without, DNA hybridization catalysis, the complementary strands continue to freely diffuse in a solution, without binding to each other.

1.1.4. Nucleic acid hybridization

Hybridization or preferential binding of a DNA/RNA sequence to its complementary sequence forms the basis for information transfer in biological systems. The resulting sequence can be a DNA/DNA or a DNA/RNA hybrid. Nucleic acid hybridization forms the basic and important foundation for many important gene expression profiling techniques (Pozhitkov et al., 2008). Figure 1.4 shows the nucleic acid hybridization of a DNA.

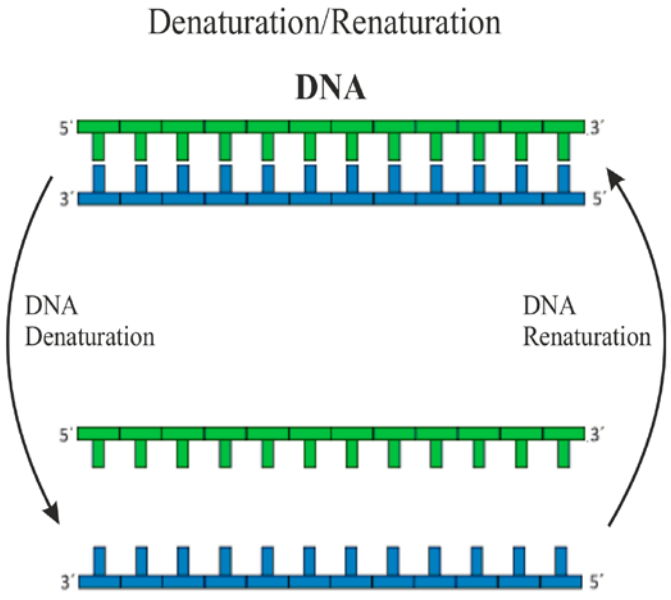


Figure 1.4: Nucleic acid hybridization

Hybridization of an ssDNA sequence to its complementary sequence under controlled conditions is necessary for the analysis and quantification of gene expression. The sequences (probes) are immobilized to a surface which can be surface or a membrane. Then the genes which are to be analyzed are hybridized to the probes in a solution. Labeling the sequences can be performed before or after hybridization. The resulting signal is detected with fluorescent or chemiluminescent techniques and analyzed for information transfer.

1.1.5. Gene Expression profiling

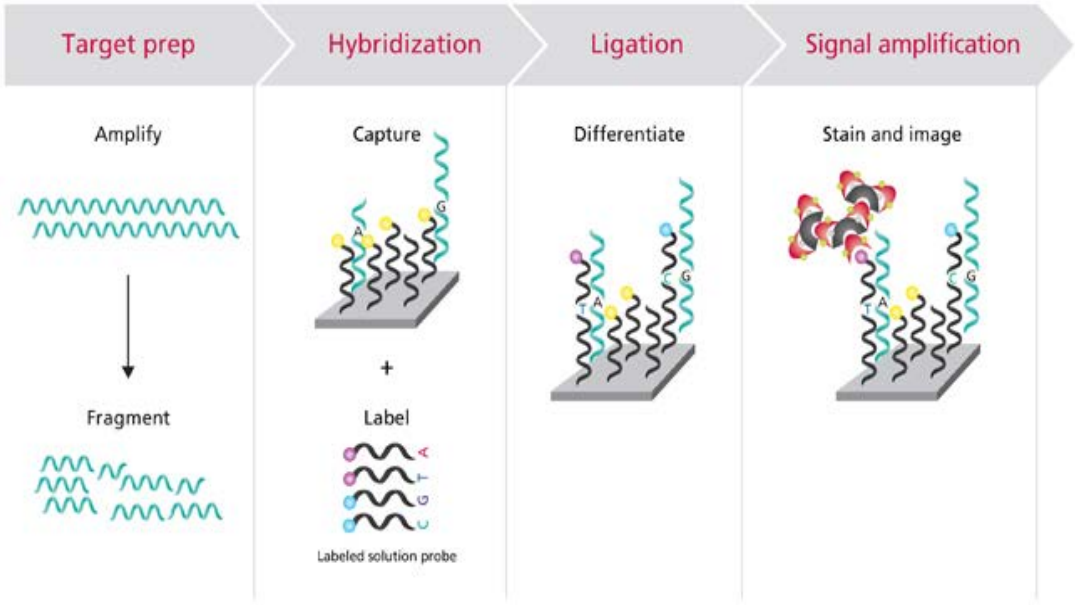


Figure 1.5: Gene expression profiling based on Affymetrix Microarray platform

Gene expression profiling is used to measure the activity of large number of genes in a single experiment. This provides a lot of information about the activity and status of multiple cells (with specific genes) at a particular instant (Schena et al., 1995). Thus the behavior or response of stress to specific conditions can be measured simultaneously. cDNA/oligonucleotide microarrays are used to investigate thousands of nucleic acid

sequences at a particular instant in an single experiment (Tarca et al., 2006). Figure 1.5 shows the gene expression profiling with Affymetrix microarray platform. Microarrays for predefined applications and microarray platforms are available commercially. Quantitative PCR (Petersen et al., 2007) and membrane based blotting techniques (Southern, 1992) are also used for gene expression profiling.

1.1.6. Problems with gene expression profiling

Gene expression profiling provides huge amounts of data about the expression/activity of genes. The sample preparation for data retrieval is mostly based on conventional random fragmentation techniques in case of microarrays and surface based hybridization techniques. PCR based specific amplification is applied in the case of quantitative PCR. With conventional techniques, the sample in question is randomly fragmented at various recognition sites using commercially available thermostable enzymes (Fukano and Suzuki, 2009). The huge amounts of data include incomplete information, lost information and unspecific information. This leads to noise generation during sample preparation and gene expression profiling. The potential of DNA microarrays is enormous. Their application is limited due the combination of various factors like dependence on statistical analysis and absence of a standard for various commercial platforms (Tan et al., 2003). Most of the commercial microarray platforms apply various statistical models (Sifakis et al., 2012). Due to this, the information generated for a sample source can vary with change of commercial platform.

Also, existing sample preparation techniques result in generation of DNA/RNA sequences which are either longer than 150 nucleotides or shorter sequences in increased amounts. If

the information source is highly fragmented for shorter sequences, there is a higher risk of incomplete information and irreversible loss of information. Longer sequences form loops an hairpin for more stability (Kuznetsov et al., 2008). This results in loss of information.

In this dissertation, we propose a novel nucleic acid retrieval technique from an information source. This information source has specific information along with sources for noise. Our technique produces specific data in reduced amounts. This results in efficient retrieval of specific information.

1.2. Previous research and shortcomings

In this section, the advances in molecular biology and various fields that lead to evolution and improvements in analysis of information flow in biological systems is presented. At the end of this section, the need for an alternative approaches is briefly explained as follows.

1. Thermostable polymerases and novel enzymes

Advances in production of thermostable enzymes/polymerases have led to developments in sequencing and amplification of difficult templates (Pavlov et al., 2004), (Stenlund et al., 1980). These restriction enzymes evolved from ones which cut at random to genetically manufactured enzymes that cut at specific recognition sites. They are popularly known as class four restriction enzymes, for example BgaI enzyme (Saupe et al., 1998). This led to fragmentation/amplification techniques that generate large amounts of data.

2. Conversion of long DNA to short fragments

Long DNA sequences, which can be either a whole genome or cDNA sequences that are extracted from a tissue sample with tumour, needed to be analysed and compared to healthy tissues. This is accomplished by generating a library (data set) of information sequences

through enzymes like DnaseI (Fukano and Suzuki 2009).

3. Evolution and advances in sequencing techniques

Initially, the source of information is fragmented and sorted into short fragments which may be random in size and length distribution. Then these fragments are sequenced. Sequencing means finding out the nucleotide composition on the DNA sequences (Weissman 1979). This uses the preferential binding of DNA as a foundation for unlocking the nucleotide composition.

Previous research has shown the importance of bioinformatic analysis for analysing huge amounts of raw data. However, it cannot be neglected that these huge amounts of data also lead to lot of problems (Greiner and Day 2004), (Lindow et al., 2012). Sometimes, these huge amounts of data may not contain the specific information needed for particular application.

Previous research has provided valuable insights into various factors that influence information transfer in biological systems like mutations, single nucleotide polymorphisms (SNP) and external factors like stress and radiation effects on cells.

These problems with understanding and monitoring information flow from DNA to proteins have been explained through various theoretical and experimental models. These models mostly depend on simulations, mathematical modelling and bioinformatic tools like molecular models (Riva et al., 2005).

There are mostly two types of experimental models

1. Comparison between wild type (healthy organisms/cells) and affected organisms/cells to study discrepancies in gene expression.
2. Experimental models that predominantly use various sample preparation techniques are

described above to generate information need for further analysis.

Most of these experiments/models are performed either to prove a hypothesis to be accurate or to generate information and then propose a hypothesis. Hypothesis is an assumption about the state of a particular gene or a cluster of genes in a biological system. The experiment is performed based on the above mentioned procedures. Samples are analysed/sequence and the hypothesis is proved either to be right or wrong.

After the experiments are performed, the data is analysed with high throughput application like microarrays. These data sets are analysed with various statistical models like Monte Carlo simulations (Draminski et al., 2008) and boot-strapping.

When a particular test is selected, particular gene sets are identified. This introduces discrepancies in data analysis between various commercial platforms as the bioinformatic tool-sets used for each platform have their dedicated specifications and assumptions. The presence of huge amounts of data places an enormous strain on the ability of these techniques. If a same set of genes are analysed multiple times on different platforms the resulting data is same, but the different statistical approaches make the comparisons incoherent (Tan et al., 2003).

Above mentioned advances in information analysis in biological systems combined with bioinformatic tools has opened possibilities for large data retrieval and analysis. This is made possible through combination of fragmentation techniques for whole genome/cDNA analysis with high throughput applications like microarrays and sequencing (Nordstrom et al., 2002; Sanger et al., 1977).

Most of the techniques used for the information transfer analysis tend to prefer ***brute-force*** method. They produce randomly large amounts of fragments from any given information source of interest. This has the following problems:

1. Large amounts of unspecific data ***cloud*** specific data
2. A fragmented mixture is generated which has the following components
3. There is a possibility of irreversible loss of information due to random fragmentation

Fragmented mixture=Specific information+Unspecific information+Mixed information-Lost information

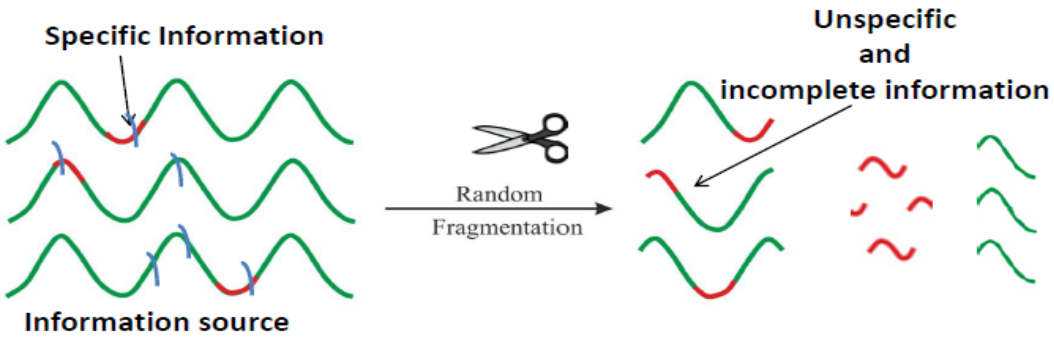


Figure 1.6: Sample preparation for gene expression profiling

In this dissertation, we propose a model to solve the fundamental problems that hamper the enormous potential of gene expression profiling. Figure 1.1 shows the problems arising from existing sample preparation methods for gene expression profiling.

When analysing information flow from DNA to active protein it is ideal to produce information that is specific to the particular stretch of DNA. This can be either a gene or a cluster of genes. By concentrating on specific regions and generating multiple shorter ssDNA sequences which have a predefined length and nucleotide composition, the following discrepancies are avoided.

1. Longer fragments form loops and hairpin structures. So part the sequences do not hybridize with their complementary sequences and result in information loss.
2. Generating reduced amounts of raw data for analysis eases the significant strain on the capabilities of microarray analysis and reduces platform dependency.
3. By generating an information mixture, dead-data which is data that has no use for the

particular experiment or hypothesis is produced thereby giving unspecific signals and information which clouds the specific information.

4. Dynamic equilibrium is difficult to obtain when more sequences compete to hybridize with the sequences on microarrays (Bhanot et al., 2003).
5. When a portion of dead data has some stretch of ssDNA that is similar to the immobilized sequences, the possibility of unspecific or partial hybridization is increased.
6. The major constraint of high throughput applications is the cost aspect. This can be reduced by fabricating microarrays with specific probes instead of large amounts of sequences, which may or may not be important for the particular experiment/hypothesis.

2. Experimental model based approach

2.1. Model for holistic approach

In this dissertation, a holistic approach is applied to build a model system that represents a complex biological system at a fundamental level. This fundamental state is *a **simple biomolecular model system*** also referred to as a ***Toy system***. It has specific information along with noise. Then a reductionist approach is applied to extract the specific information from the model system. This combination of holistic and deterministic approaches enables to understand the dynamics of information transfer in complex biological system.

Application of holistic approach specific information is systematically embedded into a noise environment that provides many possibilities for incorrect information transfer are

1. Unspecific hybridization.
2. Production of large amounts of dead data.
3. Synthesis of longer sequences which form loops.
4. Synthesis of a fragmented mixture that has dead data, incomplete information through existing fragmentation techniques.
5. Strain on high throughput applications though large data processing requirements.
6. Possible introduction of intrinsic and extrinsic noise sources and random mutations through introduction of bacterial culture as the final step to express our specific information inside a cell.

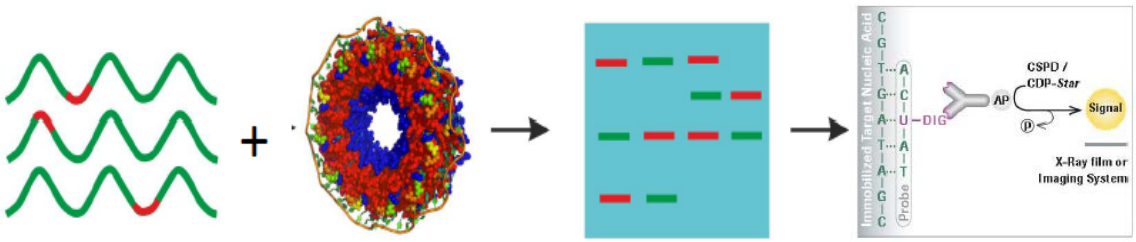
2.2. Model for reductionist approach

By using a reductionist approach for retrieval of specific information, the possibility of incomplete and defective information retrieval is introduced. All the noise sources that were

introduced through holistic approach can influence the identification, isolation and retrieval of accurate information from the Toy system. Thus, the shortcomings of existing information retrieval and analysis techniques are understood.

This Toy system is a complex DNA mixture which has noise and specific information that is realized through various molecular biological and biochemical techniques. From this toy system reduced amounts of information which is highly specific in nature is retrieved. This model depends on the biomolecular interactions that govern the information transfer in various processes like molecular binding, DNA replication and hybridization.

The intention behind choosing a systematic model based approach is to replicate the possibilities of noise sources and loss of specific information. Instead of comparing one set of data to another and proving a hypothesis, this Toy system introduces most possible noise sources to cloud the reduced amount of specific information. Successful retrieval of this specific information depends on the ability to remove the noise sources.



Specific Information (Red) Noise Retrieval Verification

Figure 2.1: Model for information transfer in a Toy system

The novel MEA method used to identify and amplify specific ssDNA sequences that represent specific information provides a solution to separate noise sources during the initial step of information retrieval. This method functions through the control of fundamental biomolecular interactions like DNA-DNA hybridization for information transfer instead of

random fragmentation. The retrieved data is screened through various molecular biology and biochemistry based techniques like gel electrophoresis and HPLC. Then the information is purified for removal of possible noise sources and subsequently verified through blotting for successful information transfer and further through Sanger sequencing for verifying the nucleotide composition. This can significantly reduce the problems with conventional techniques by generating reduced amounts of short specific information sequences.

3. Methods

Keywords: PCR, gel electrophoresis, cloning, microarrays, blotting, sequencing

- **PCR:**

  (Polymerase Chain Reaction) is used to exponentially amplify a specific region or whole DNA template to create multiple copies of amplified DNA product. A standard PCR reaction uses the following materials.

- **Nucleotides:**

  Is the fundamental building blocks of any DNA sequence, they consist of the four bases adenine, thymine, cytosine and guanine (A, T, C, G; in RNA thymine is replaced by uracil [U]), a sugar and at least one phosphate group; without the phosphate group these building blocks are known as nucleotides.

- **Primer:**

  It is a short DNA fragment with a defined sequence that functions as starting point for polymerases.

- **Polymerases:**

  Are the enzymes that link individual nucleotides together to form long DNA or RNA chains.

- **Hybridization:**

  Is the (annealing) joining of two complementary DNA (or RNA) strands to form a double strand.

- **Complementary DNA:**

  Ability of two strands of DNA/RNA to bind each other and forms a double strand of all

perfect pairs. This property plays an important role in transfer of information through DNA sequence.

3.1. PCR principle

The PCR is commonly carried out in a reaction volume of 10–200 uL in small reaction tubes (0.2–0.5 uL volumes) in a simple thermal cycler. The reaction tubes are subjected to heating and cooling to the temperatures that are prescribed in the protocol. Annealing temperature is generally dependent on the primer melting temperature.

Peltier effect is used in all conventional thermo cyclers. By a reversal of electric current, a sample holding block (which holds the PCR reaction tubes) is heated and cooled to pre-set temperatures. The temperatures used and the duration they are applied in each cycle depend on a variety of parameters.

These include the thermostable polymerases/enzymes, the divalent ion concentration and dNTPs in the reaction, length of the template and finally the primer melting temperature (Tm) (Abalaka and Henry 2011).

The stepwise methodology of a general PCR reaction is illustrated in Figure 3.1.

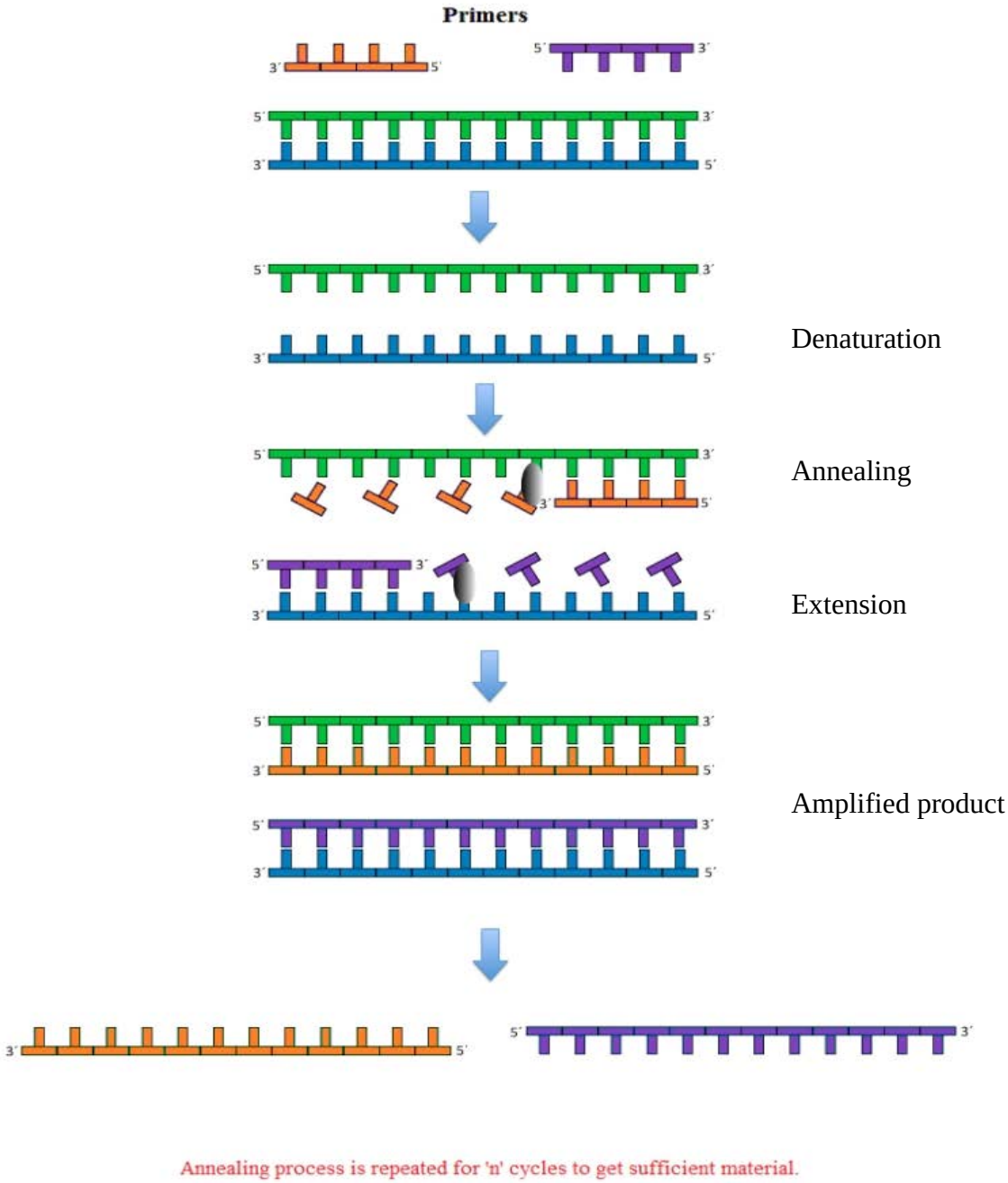


Figure 3.1: Methodology of PCR.

The first step in the process is to extract the DNA template from the sample. If it is a tissue or blood sample, then DNA is extracted either directly or indirectly using various commercially available kits and molecular biology techniques:

- **First Step-Denaturation:**

  Temperature rise is about 90-95^{O}C. This denatures double stranded DNA into single strands.

- **Second Step-Primer Annealing:**

  Temperature is then reduced to about 50-65^{O}C. This allows the two primers anneal, at specific points on the denatured single-stranded DNA of the target sequence.

- **Third Step-Extension:**

  The temperature is raised to 70-74^{O}C and the DNA-polymerase enzyme catalyses the duplication/copying of the target sequence. This step uses the free nucleotides as building blocks and starts at the annealed primer regions on each single strand.

This results in two double-stranded DNA fragments, which are duplicates of the original target sequence. The temperature cycling from first to third step is then repeated 30-40 times, creating an exponential increase in the copies of the target sequence at each cycle. This produces sufficient amplified DNA for reliable detection from single and minute amounts of target sequence in 2-4 hours. The temperatures used and the duration they are applied in each cycle depend on a variety of parameters. These include sample handling, the polymerases used for DNA amplification, the divalent ion concentration and dNTPs in the reaction, and the melting temperature (Tm) of the primers and finally the length of the template.

3.1.1. PCR stages

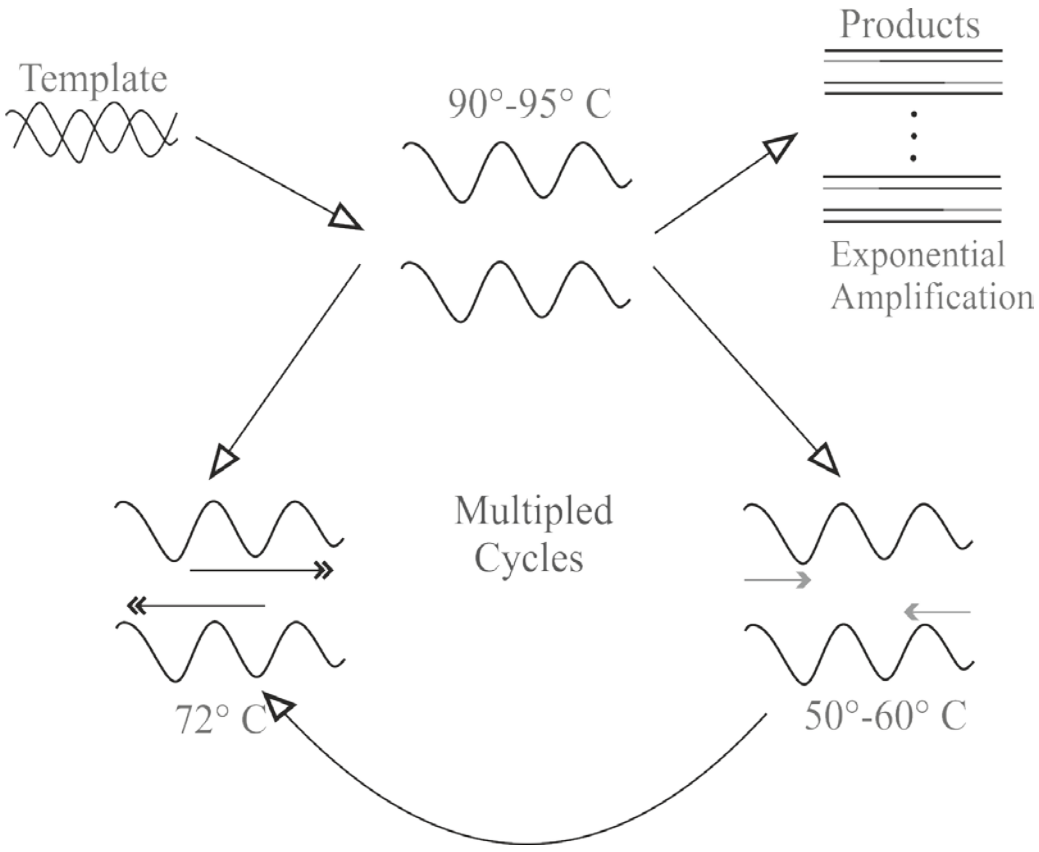


Figure 3.2: PCR temperature cycles

The PCR process consists of three steps:

1. **Exponential amplification:**

   At every cycle, the quantity of amplified/duplicated product is doubled (100% reaction efficiency is assumed).

2. **Levelling off stage:**

   DNA polymerase becomes inactive with the progress of the reaction. Also, consumption of reagents such as dNTPs and primers causes them to become limiting.

3. **Plateau:**

   All limiting materials that are necessary for amplification like dNTPs, enzymes and primers are used up by this stage. So there is no more amplification of the template.

3.2. Variations of PCR

Modifications of PCR that are used in this dissertation

3.2.1. Qualitative PCR

This modification of PCR is used only for detecting a specific DNA segment, for example a single gene in a genome. Qualitative PCR is an extremely sensitive.

3.2.2. Singleplex and Multiplex PCR

Singleplex PCR uses only one pair of primers. Multiplex PCR uses multiple templates and several primer sets in the same reaction tube (Wen and Zhang, 2012). Presence of multiple primers may lead to mispriming with other primer pairs. So, selection of unique primer pairs is an important parameter in multiplex PCR.

3.2.3. Linear-After-The-Exponential (LATE)–PCR

It uses two primers like conventional PCR. However, one of the primers is in excess concentration and the other primer in limiting concentration. After initial exponential amplification with both primers, the limiting primer is diminished. So, the resulting product is dependent only on the excess primer and is linearly amplified. This results in a dsDNA amplified product and an ssDNA linearly amplified product.

3.2.4. Linear PCR

It uses single primer instead of a primer pair. This results in linear amplification of the template instead of exponential amplification. This technique is of great importance in our work and will be described in detail (Sanchez et al., 2004).

3.2.5. Colony PCR

Colony PCR is used after a transformation to screen colonies for the desired plasmid. Primers are used to amplify a PCR product of known size. Thus, colonies which give rise to an amplification product of the expected size are considered as positive colonies with intended DNA sequence.

3.2.6. Primer extension analysis

It is used to specifically locate the position of 5-end of RNA. Here, RNA is used as a starting template. An end-labeled oligonucleotide (primer) hybridizes to the template RNA. Reverse transcriptase extends the primer along the template in the presence of deoxynucleotides. Finally, RNA undergoes reverse transcription to cDNA. Further analysis with denaturing Polyacrylamide gel can accurately quantify the amount of transcribed product. The length of the reverse transcribed cDNA shows the distance between the primer and the 5´-end of the RNA (Syvänen, 1999).

In this dissertation, the accuracy of qualitative PCR is merged with simplicity of primer extension to develop a novel technique for accurate information transfer and verification in a complex DNA mixture.

3.3. Modified End Amplification Technique- MEA

MEA technique uses results in linear amplification. This results in singular copies of target sequence. This means one copy of ssDNA sequence from one cycle. It differs from linear PCR techniques in the following ways.

- The single primer used in this technique extends till the end of the template. So there is no scope for further extension.
- This results in linearly amplified ssDNA sequences, which have a specific length.
- By modifying the primer binding region on the DNA template, the length of the linearly amplified sequences can be progressively influenced.
- Also, multiple primers can be used to amplify multiple regions on the two strands of the DNA template.
- Single primer
- Increased number of cycles
- Adding fresh Taq polymerase after 50-60 cycles (optional)
- No end amplification at $72^{o}C$

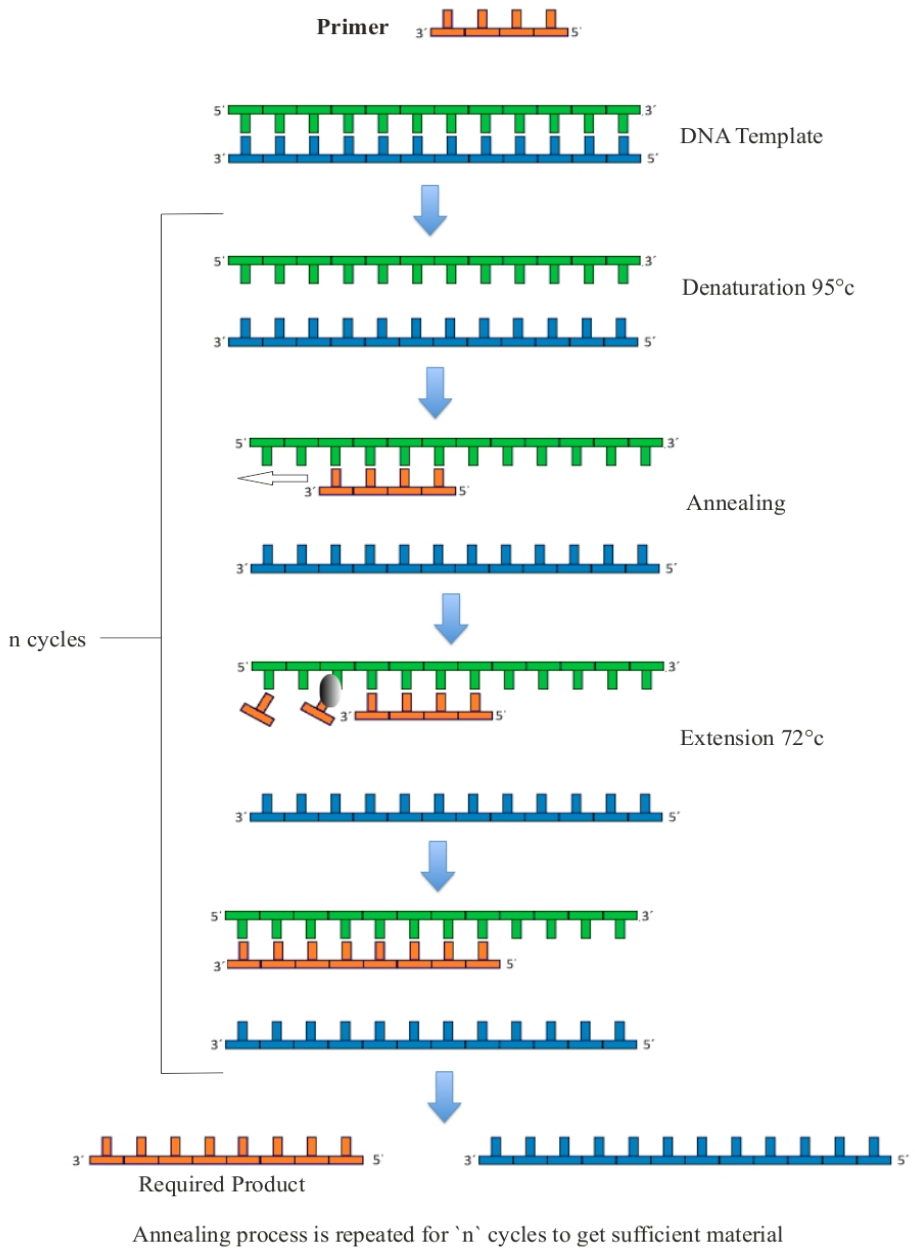


Figure 3.3: Modified end amplification technique

The protocol used for MEA technique, has following modifications to stand PCR protocol. This end amplification technique was tested with Bca polymerase which is suitable for isothermal amplification. In isothermal amplification, there are no multiple cycles for amplification (Vincent et al.,, 2004). The amount of amplified product depends on template and primer concentration. Isothermal amplification eliminates the need of thermo cyclers.

3.4. Gel electrophoresis

Electrophoresis is one of the most widely used techniques in every molecular biology laboratory.

- Agarose
- PAA-Urea gel

Electrophoresis is a procedure in which biomolecules such as DNA, RNA and Proteins are separated based on their molecular weight and charge. An electric field is applied to a gel matrix which is a Carbohydrate complex. This forces various biomolecules through the gel matrix. As the biomolecules are forced to move through the gel, they encounter resistance which slows down their rates of migration. With Gel electrophoresis, smaller biomolecules have a faster migration rate than larger ones. Migration distance in the gel is used to determine the size/length of biomolecules. This is accomplished by comparing the biomolecules in question with a reference mixture of biomolecules on the same gel. The **phosphodiester** bonding between the **complementary nucleotides** (A-T, C-G) provide a negative charge to the DNA. When an electrical field is applied to the gel, the negatively charged DNA migrates towards the positive electrode.

Importance application of gel electrophoresis is in restriction digestion of DNA. Restriction enzymes cut/slice the larger DNA sequence into several shorter sequences. When this mixture of shorter sequences is run in a gel, the sequences are separated according to their size/length. After electrophoresis, the gel is soaked in a solution containing a fluorescent tag. This tag, generally Ethidium Bromide intercalates between nucleotide bonds of DNA and results in intense fluorescence when the gel is exposed to UV illumination. Ethidium Bromide is a mutagen, but is preferred to various safe dyes like SYBR green and DAPI as it is considerably less expensive.

3.5. Horizontal gel electrophoresis

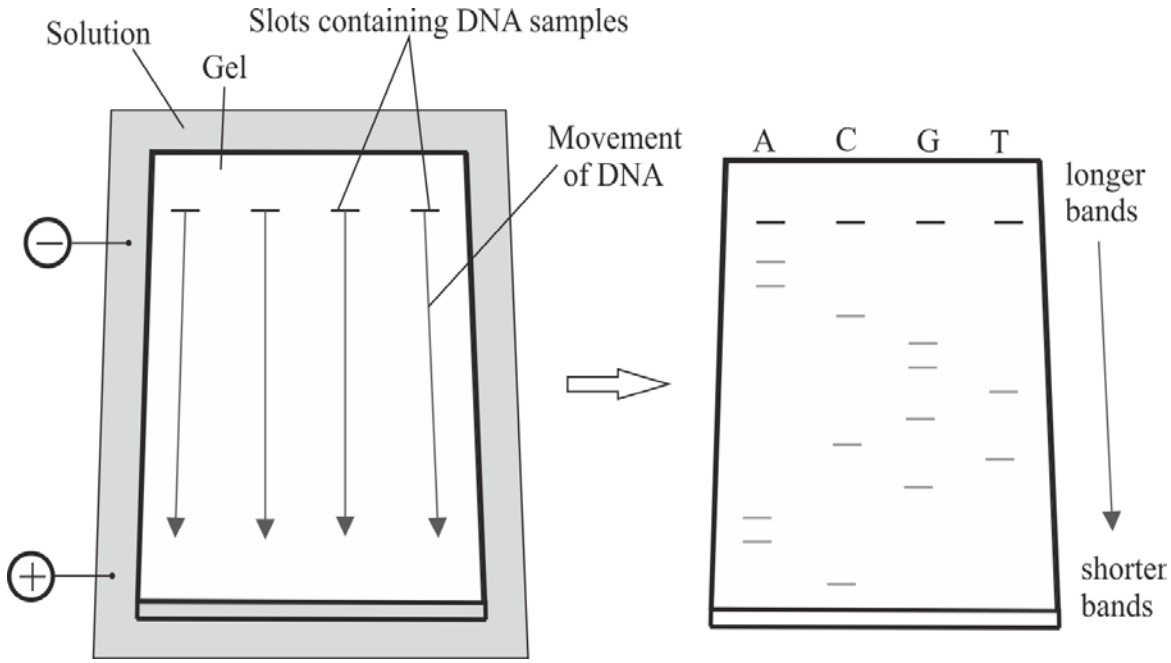


Figure 3.4: Agarose gel electrophoresis

Horizontal gel electrophoresis is used for analysing/separating larger nucleic acids like DNA and RNA. For this purpose purified agarose is used to form the gel. Melted agarose after cooling forms a matrix with large pores. When electric field is applied to the agarose gel, the nucleic acids are pushed through the pores and size based separation occurs. agarose gel electrophoresis chambers are horizontal.

3.6. Vertical gel electrophoresis

It is used to separate smaller nucleic acids like ssDNA, RNA and proteins. Polyacrylamide/Bis acrylamide mixture is used to form a gel matrix together with cross linking agents like APS and TEMED. To separate proteins, SDS-PAGE gels are used and for

ssDNA/RNA denaturing gels are used.

In this dissertation, only Denaturing PAA-Urea gels are used to separate ssDNA. Urea acts as a denaturing agent and removes the secondary bonds between DNA, to keep the ssDNA linear without any loops. Acrylamide-Bisacrylamide powder is used in a 10-12% end concentration. This is necessary to separate ssDNA sequences between 20-100 nt.

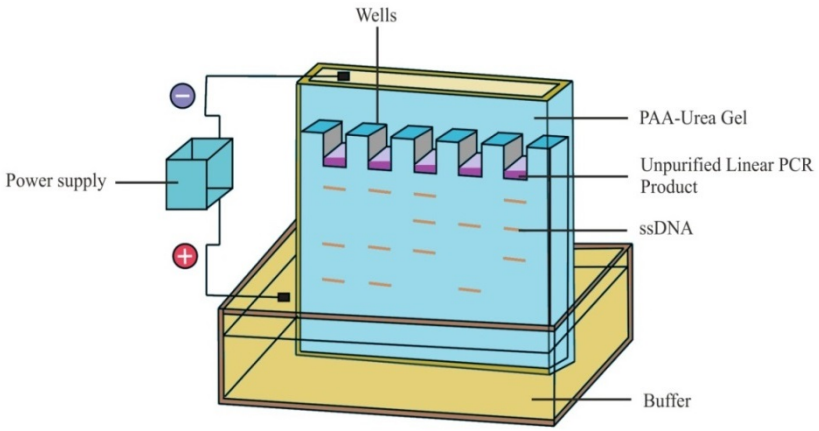


Figure 3.5: Vertical gel electrophoresis for ssDNA separation

Acrylamide powder is a potential Neurotoxin, so care should be taken when handling these gels before they polymerise. Also, Acrylamide does not polymerize in presence of O_2. So the gel solution is poured between these glass plates, to avoid contact with atmosphere.

As ssDNA is linear in denatured gels, it is difficult to use Ethidium Bromide for staining these gels. So silver staining is used to visualize ssDNA/oligonucleotides in the gel. Silver is highly sensitive to nucleic acids. All the protocols regarding gel preparation and staining are given in the Protocols section.

3.7. Nucleic acid purification techniques

Various commercially available columns were used to purify the products from both conventional and our modified MEA techniques. After PCR or any nucleic acid amplification techniques, the obtained product should be purified to remove the excess primers, enzymes and other possible inhibitors for down-stream applications.

In this dissertation dsDNA was obtained from Singleplex PCR, colony PCR and Plasmid DNA after bacterial transformation. For these purposes, commercially available gel extraction and plasmid purification kits were used. For purifying ssDNA from our modified MEA techniques, biotin-streptavidin magnetic beads and Polyacrylamide gel extraction kits were used. Biotin-streptavidin bond is the strongest non-covalent interaction between protein and a molecule.

The primers that generate MEA products (ssDNA 40-100 nt) are labelled with Biotin on the 5-end and purchased through Metabion GmbH. Later MEA technique is used and ssDNA products are obtained. Then Streptavidin coated magnetic beads are added to the solution. Later MEA products with Biotin label bind to the magnetic beads. The bound ssDNA are separated with a magnet and the Biotin-Streptavidin bond is broken with harsh denaturation conditions using formamide and higher temperatures like 90 ^{O}C.

Here, modified version of above technique is explained. Initially primers that amplify the dsDNA sequence (template) are purchased with a Biotin label on the 5´-end from Metabion GmbH. Then Singleplex PCR is performed to produce a dsDNA template with Biotin label on both 5-ends. Later MEA technique is used to linearly amplify ssDNA from the Biotin labelled template. Now the unpurified MEA product solution is made alkaline. This results

in dissociation of the dsDNA template to ssDNA sequences.

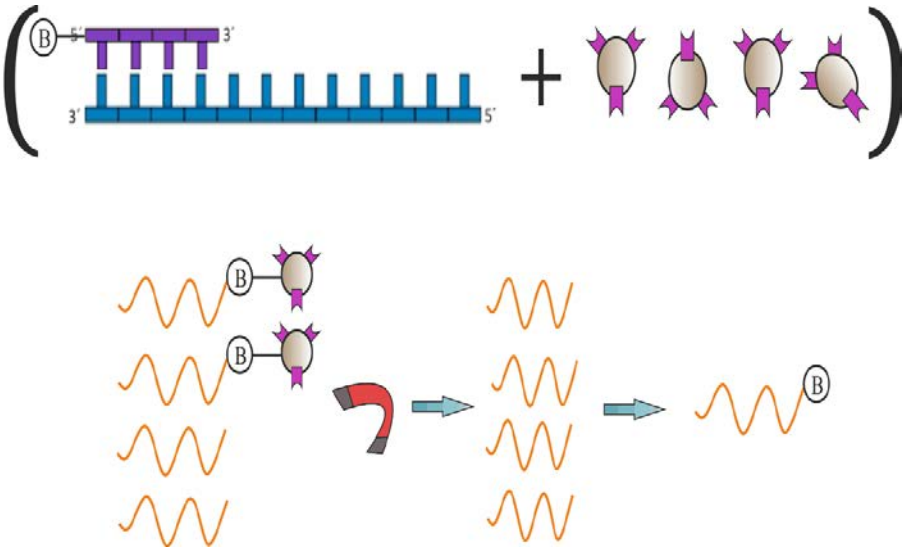


Figure 3.6: Biotin-Streptavidin nucleic acid purification

Then Streptavidin coated magnetic beads from Promega GmbH were added to the unpurified MEA product solution. So the Biotin labelled ssDNA sequences from the template bind to the magnetic beads. They are separated with a magnet. Then the unlabelled ssDNA sequences (MEA technique products) are eluted.

3.8. High-performance liquid chromatography

Also known in some cases as **high-pressure liquid chromatography**, HPLC is a chromatographic technique used to separate a mixture of compounds to identify, quantify and purify the individual compounds in the sample of interest.

In HPLC, there are two kinds of phases are used. It depends on the column material characteristics.

3.8.1. Normal Phase

If the environment is non-polar, then hydrophilic molecules bind with each other. Initially, the analyte is mixed with less polar mobile phase and injected into the column. The stationary phase-column binding material is made to be hydrophilic. The materials A, B and C are absorbed to the particles in the column. If the polarity of the mobile phase is increased, the absorption on analytes to the column decreases. Finally, the molecules in the analyte break free and are eluted from the column along with the mobile phase. Depending on the polarity of the materials in the analyte-sample mixture, the retention time of subsequent compounds in the analyte changes. In the Figure 3.7, analyte has compounds A, B and C. Compound A has less polar, so it breaks free from the column first, followed by B and C. The compounds pass through a detector and can be collected using fractionators for analytical analysis. These compounds are separated based on the retention times are shown in a chromatogram.

3.8.2. Reversed phase HPLC (RP-HPLC)

In Reverse phase HPLC; stationary phase which is the column material is non-polar. The mobile phase, which is aqueous, is polar when compared to the stationary phase. When the polarity of the mobile phase is decreased, the compounds which are more polar elute faster. The retention times for compounds with less polarity are higher.

Surface modified Silica (modified with Rme_2SiCl) is a common stationary phase. Here R is a straight chain alkyl group like $C_{18}H_{37}$-known as C_{18} columns or C_8H_{17}-C_8 columns. With C18/C8 columns, less polar molecules have longer retention times.

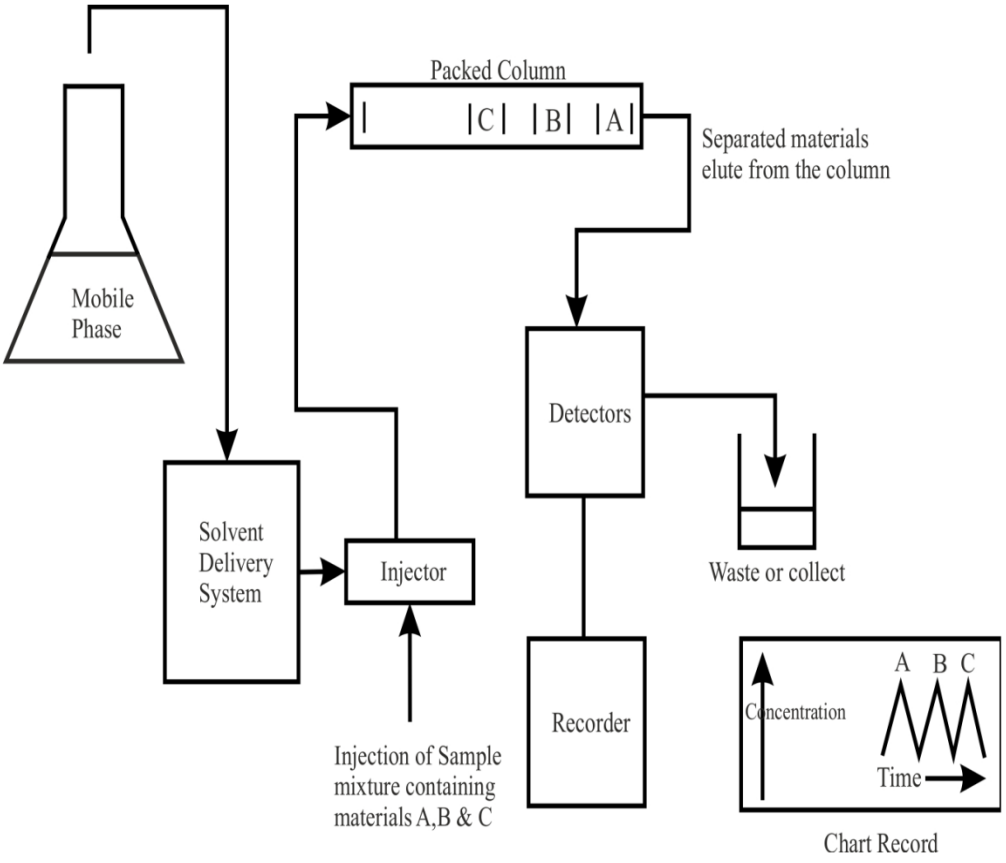


Figure 3.7: HPLC methodology

During these experiments HPLC Dionex Summit System consists of Pump P680, Auto sampler ASI-100. Chromeleon 6.8 software was used for evaluation and analysis was used with a C_{18} Xbridge column from Waters Corporation. Detailed protocols are provided in the protocols section.

3.9. Molecular cloning

Cloning uses a single copy of DNA sequence (sample) from a living cell and creates multiple number of cells with the same DNA sequence. The copies have the same nucleotide composition as the sample and with similar genetic properties.

Our interest is in molecular cloning. This is recombinant DNA technology that uses two different DNA sequences. One of them is the target-DNA sequence from sample. Second is the vehicle/vector that transports the target sequence into a host organism. In molecular cloning techniques, this host is generally modified *E. Coli*.

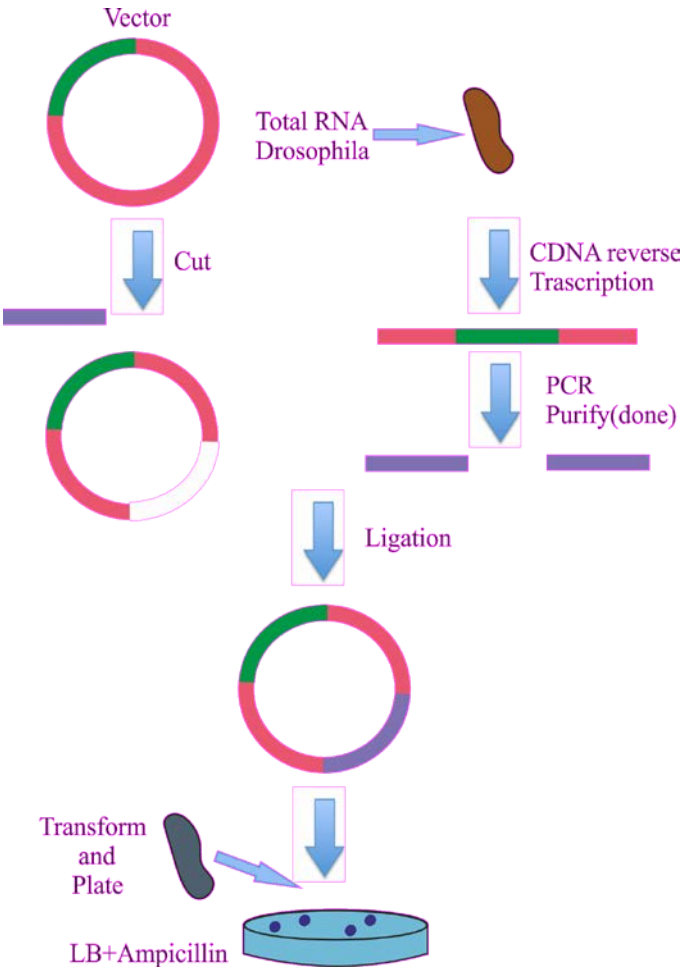


Figure 3.8: Molecular cloning

Vectors or transport vehicles are generally modified E.Coli plasmids of maximum 3 kb in length that replicate in the host cells and produce large amounts of cloned/copied plasmids. For the purpose of cloning, the vectors are cut at a single region and a linearised vector is generated.

Compositions of a modified E.Coli plasmid vector are

- Important nucleotide sequences required for cloning purposes
- Origin for replication
- Gene for specific drug resistance. For example, Ampicilline, Kanamycin etc.
- A region (multiple cloning site) for insertion of target DNA sequences for cloning

After ligation of the target DNA sequence and the linear vector/plasmid, a circularized recombinant plasmid is generated. This plasmid is incubated with Host cells (modified E.Coli cells) under conditions that allow transformation of the plasmid into bacterial cells.

Transformation conditions are:

- Either electroporation or heatshock for the recombinant plasmid to be transformed into E.Coli cells.
- Appropriate antibiotic that is selective to the cells.
- Culture plates (LB-Agar plates with selective antibiotic)
- Incubation of culture plates at 37^0C for overnight.

For every individual cell that is transformed with an identical plasmid, a colony results on the culture plates. All these colonies have identical plasmids with the target DNA sequence inside. Now, an individual colony is further cultured in appropriate culture medium with selected antibiotic. After extraction of plasmid with various Plasmid isolation kits, plasmid containing the target DNA of interest in obtained in very high concentrations (100-500 ug/uL).

This DNA cloning is very popular in molecular biology applications for purifying and

multiplying target DNA sequences from a complex DNA mixture.

3.10. Blotting

Blotting is used to detect DNA/RNA/Protein-biomolecules from a complex mixture of biomolecules. If the blotting technique is used to detect specific DNA sequences by hybridization with its complementary sequence, then it is referred to as Southern blotting. It was named after its inventor, Edward Southern (Southern, 1992).

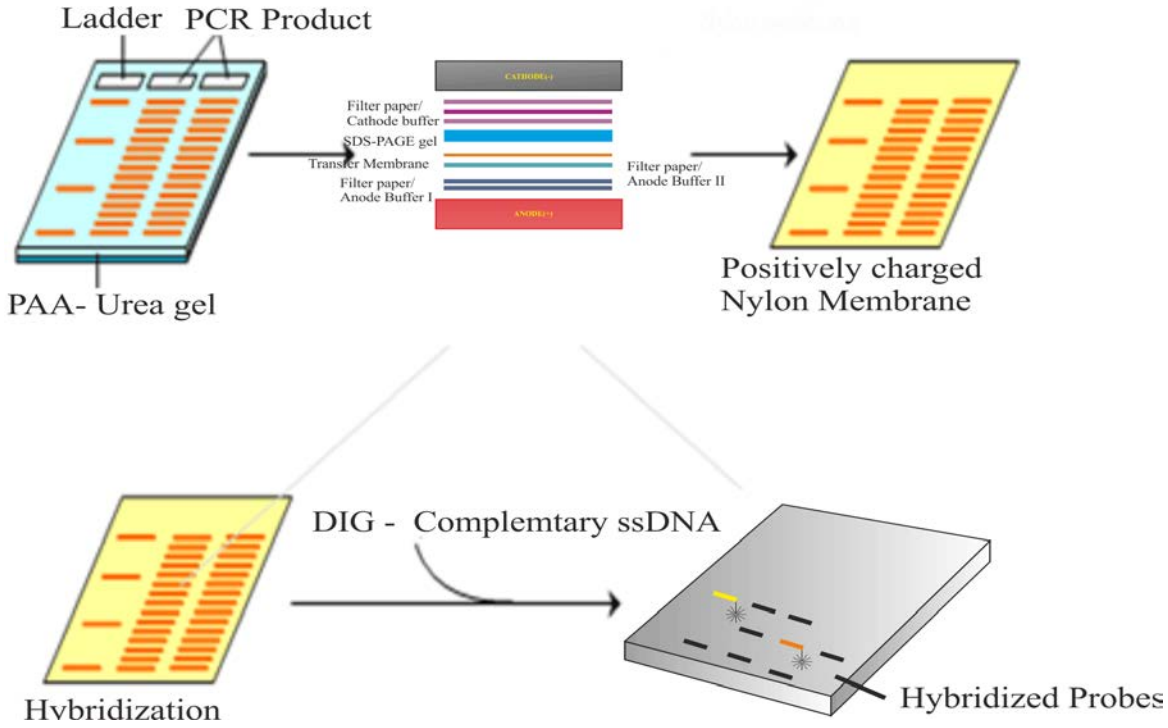


Figure 3.9: Southern blotting with Hybridization.

In conventional Southern blotting, a mixture of various DNA fragments is prepared through various procedures like random fragmentation or restriction digestion. These fragments are separated on either agarose or Polyacrylamide gels through gel electrophoresis.

For our specific Southern blots, ssDNA sequences prepared through linear PCR or our novel MEA methods are separated on 12% denaturing PAA-Urea gels. These ssDNA sequences (Probes) are printed or transferred on a PVDF membrane through Semi-dry blotting technique. After the ssDNA sequences are transferred on the membrane, the ssDNA sequences are hybridized with their respective complementary ssDNA sequences (Targets). These targets are tagged with Digogxigenin on the 5´-end. Later, anti-DIG-Conjugate is bound to the DIG-labelled ssDNA sequences. These conjugates exhibit Alkaline Phosphatase activity. This is detected with various techniques like colour or chemiluminescent detection (Lanzillo, 1991). The protocol is described in detail in Protocols section.

3.11. DNA Microarrays

DNA/oligonucleotide microarrays are solid substrates with oligonucleotides (for example ssDNA sequences) of lengths between 15-100 nt immobilized at the end (preferably 3´-end) to the surfaces (Riva et al., 2005). The oligonucleotides are attached to the solid surface by various chemical modifications. The immobilization is achieved either by attachment of oligonucleotides though modification at the end of the sequences or by surface treatment to the surfaces like silanization.

If PCR products are attached to the surfaces and the length of attached sequences/PCR products is more than hundred nucleotides, then these microarrays are termed as cDNA microarrays. However, these cDNA microarrays are restricted to gene expression profiling studies, as it is difficult to distinguish between target sequences of single base modifications.

There are various methods to attach ssDNA sequences to the solid substrates

The first gene expression assay on a printed microarray was reported in 1995. They used a contact printing technique. In this technique, pre-fabricated nucleotides are mixed with a spotting solution and printed on to surface with a pin. They employed a *contact printing technique* for deposition of tiny spots. This technique is later popularized as spotting (Seliger et al., 2003). These spotting techniques are fully automated with the usage of Microarray robots (*arrayers*) for completed automated procedures.

Light-directed *in situ* synthesis approach (Schena et al., 1995) is completely opposite to spotting. The probe molecules (ssDNA oligonucleotides) are fabricated *in situ*, i.e. nucleotide on nucleotide.

- Ink-jet techniques (based on piezoelectric deposition) are used for *in situ* synthesis of microarrays (by deposition of phosphoramidites) and also for *spotting* of pre-synthesized DNA.
- A rather novel technique is the electrochemical *in situ* synthesis of DNA microarrays.

In this dissertation oligonucleotide microarrays are fabricated with Light-Directed in situ synthesis method. Figure 3.10 shows the schematic of microarray synthesis and hybridization.

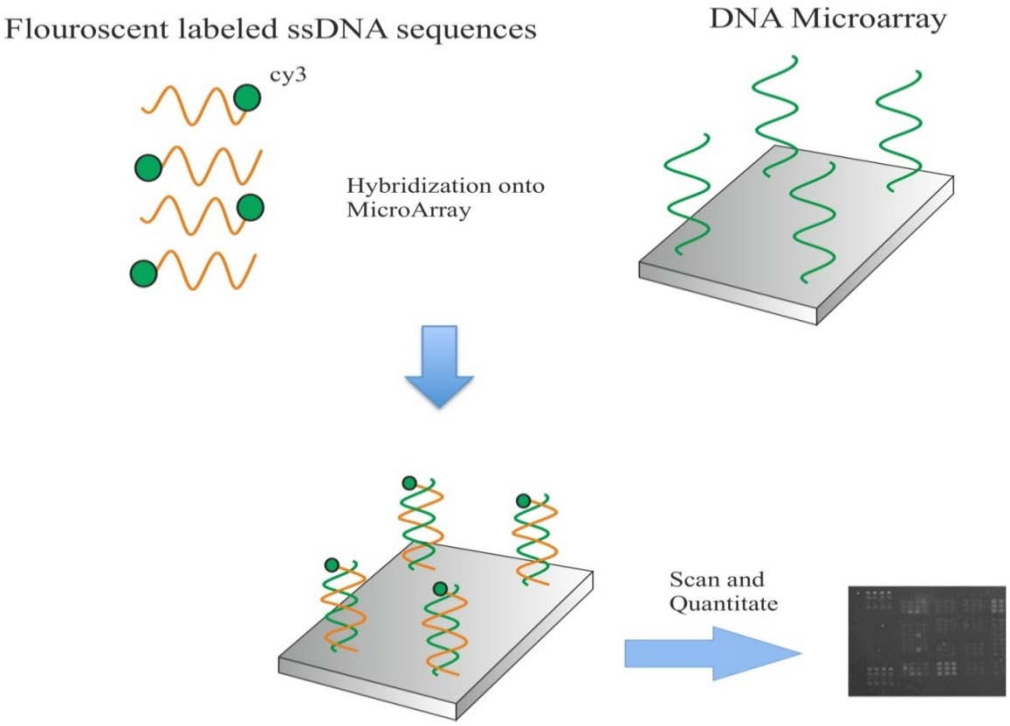


Figure 3.10: DNA Microarray methodology

All the protocols for surface preparation and hybridization are provided in the Protocols section. The microarray picture shown in this Figure was provided by Christian Trapp, Department for Biological Experimental Physics, Saarland University.

3.12. Sanger sequencing

Sanger sequencing is a technique used of reading the nucleotide/sequence composition of any DNA sequence. This works through the pre-defined insertion of chain-terminating dideoxynucleotides through DNA polymerase (Sanger and Nicklen, 1977). It was developed by Frederick Sanger and colleagues in 1977. It is one of the widely used methods for determining the sequence composition for the last 25 years. Now Next-Gen sequencing

improves Sanger sequencing through the addition of automated analysis (Mardis, 2008).

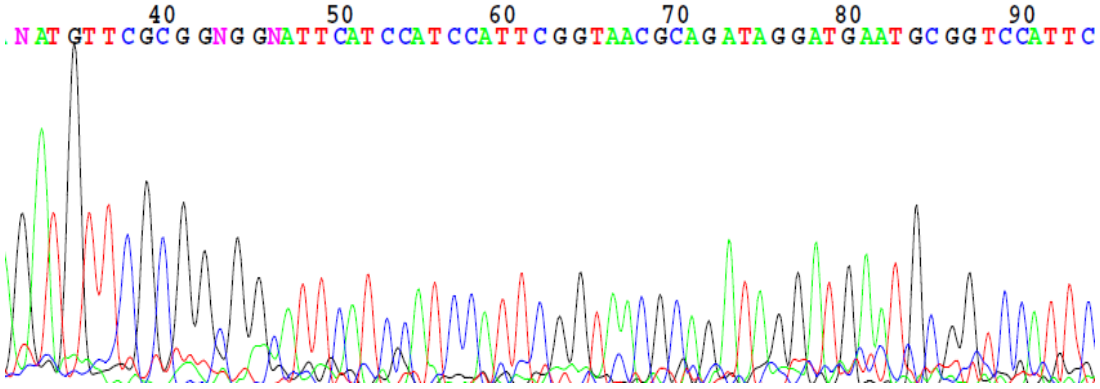


Figure 3.11: Chromatogram from Sanger sequencing

Sanger sequencing uses modified nucleotides dideoxynucleotides ddNTPs along with deoxynucleotides dNTPs to stop the extension of complementary strand on the template. The mentioned ddNTP lacks the 3´-OH group which is important for phosphodiester bonding. Thus the extension stops at specific bases. The ssDNA strand obtained after denaturation of dsDNA is used as template. Sometimes ssDNA strand can also be used as template. This is similar to linear PCR as only one primer is used. The ddNTPs are fluorescently labelled and detected through gel electrophoresis with a fluorescent detector. Every ddNTP produces a different and specific colour for the particular base. Thus the nucleotide composition is determined.

3.13. Protocols

1. PCR
2. MEA technique
3. Primer extension analysis
4. Gel electrophoresis

5. Molecular cloning and Bacterial transformations
6. Colony PCR
7. Automated Sanger sequencing
8. Southern blotting
9. DNA microarrays

- **PCR**

Template Lambda DNA from Fermentas GmbH. Cat.Nr.:- SD0011

- **Enzymes**

Taq Polymerase: Axon GmbH Cat.Nr.:- 22466

Pfu Polymerase: Genaxxon GmbH, Cat.Nr.:- M3004.0250

- **Nucleotides**

dNTP mix: 100 mM Fermentas GmbH, Cat.Nr.: R0181

ddCTP-cy3: dideoxynucleotides with cy3 labelling- Jena Bioscience GmbH, Cat.Nr.: NU-850-CY3

Primers were purchased from Metabion GmbH.

- **Primer extension analysis**

The Ladderman DNA Labeling Kit was ordered from Takara Bio Europe GmbH. This kit is typically used for primer labeling (primer extension with incorporation of labeled

nucleotides. For our purpose, instead of random primers provided with the kit, our specific primers were used to generate ssDNA of predefined lengths. This kit has Bca polymerase which enables linear amplification without need for multiple amplification cycles.

Ladderman labeling kit from TakaraBio GmbH Cat.Nr.: 6046

Primers: were ordered from Metabion GmbH. They have a concentration 100 uM in lyophilised form. They were diluted to various concentrations for subsequent downstream applications.

- **Standard PCR**

PCR was performed according to protocols provided by Axon Lab. Initial Primer concentration was 100 uM. For singleplex and multiplex PCR reactions, primer pairs (exponential amplification) were used. Here, final primer concentration was 50-75 nM.

- **MEA technique**

A single primer was used for linear applications. The templates for MEA technique were four amplicons-dsDNA sequences amplified from lambda DNA template. For all the experiments with 40 nt sequences, the following primer was used. Primer concentration was 200-300 nM. This was used, to guarantee ssDNA product needed for downstream applications.

dsDNA product (GR) from Lambda DNA genome used for ssDNA product amplification.

❖ **GR1**

- Primer for 25 nt ssDNA product TATAAGGGGATGTATGGCGA
- Primer for 45 nt ssDNA product CGGGTATCCAGCTTCTCCTT

❖ **GR2**

- Primer for30 nt ssDNA product TATAAATTCTGATTAGCCAG
- Primer for 50 nt ssDNA product ACGCCAGTCGCCACTGCCGG

❖ **GR3**

- Primer for 35 nt ssDNA product AGATGTTGAGCAAACTTATC
- Primer for 55 nt ssDNA product CAAAATTGAAATCAAATAAT

❖ **GR4**

- Primer for 40 nt ssDNA product GTTCGCGGCGGCATTCATCC
- Primer for 80 nt ssDNA product GGATGTTACGTCATAAAGCC

❖ **Gel electrophoresis**

Agarose and PAA-Urea gels were prepared through standard protocols provided in Sambrook and Maniatis. Staining for agarose gels was performed with Ethidium Bromide. Agarose gel extraction kit was used for extracting and purifying dsDNA fragments from agarose gels.

Invisorb-Gel extraction kit Cat.Nr.:- 1020300200

For PAA-Urea gels, gel cassettes from Anamed GmbH were used. For silver staining, Thermo quicksilver kit from Thermo Fisher GmbH was used. ssDNA concentrator kit from Zymo research GmbH was used to purify and concentrate ssDNA sequences. For all ssDNA gels, 20-100 oligo standards from IDT technologies GmbH, was used as reference.

Pierce color silver stain kit Cat.Nr.:- 24567
Anamed gel cassettes Cat.Nr.:- AN90010
Oligo standard IDT GmbH, Cat Nr.:- 51-05-15-02

❖ **Southern blotting**

The 40 nt product of the MEA technique was tested with southern blotting for hybridization. The 40 nt product was run on PAA-Urea 12% gel and the ssDNA sequences in the gel were blotted on to a charged nylon membrane. This southern blotting protocol was taken from the Doctoral dissertation of Jochen Meier from Department of Cell biology, Saarland University. The end concentration of the complementary sequences for hybridization was 25 ng7ml. Hybridization solution from

Roche GmbH was used to facilitate hybridization in solution.

EasyHyb solution from Roche GmbH Cat. Nr.:- 11603558001

❖ **DNA microarrays**

The protocols, materials and methods were taken from the Doctoral dissertation of Thomas Naiser, Department of experimental physics, Bayreuth University.

❖ **Molecular cloning**

Fermentas cloneJET TOPO cloning kit was used to clone PCR products. The kit has a linearized vector and enzymes for blunting and ligation. Cloning and Colony PCR were performed according to the manual provided by the manufacturer.

CloneJET TOPO cloning kit from Thermo GmbH Cat. Nr.:- K1231

❖ **Bacterial transformation**

TOP10 chemically competent cells were used for bacterial transformation. Standard heat shock protocol was used according to Maniatis and Sambrook molecular biology techniques.

❖ **Streptavidin Magnetic beads purification**

Magnetic beads were purchased from Promega GmbH. THE nucleic acid purification, denaturation and elution were performed according to Promega protocol.

Streptavidin Magnesphere beads from Promega GmbH Cat. Nr. Z5481

❖ **Automated Sanger sequencing**

The unpurified PCR products from colony PCR shown in Section 6.9 were sent to Gatc Biotech for Sanger sequencing analysis. Gatc viewer software was used to generate the Chromatogram from raw sequencing data.

4. Results

4.1. Complex DNA Mixture

In the Results section, initially specific ssDNA sequences of specific length and predefined nucleotide composition are identified from a template (lambda DNA genome). Various molecular biology and biochemistry techniques are used to isolate these sequences and embed them into a complex DNA mixture which has higher complexity and longer biomolecules. This higher complexity creates problems in information identification and transfer. This is our Toy system/biomolecular system.

Then various techniques are to identify and retrieve specific ssDNA sequences-accurate information from the model system. Our novel modified end amplification (MEA) method is used for information retrieval. This information is checked for fidelity with various hybridization techniques.

4.1.1. Realization of a complex DNA mixture

Steps for realising a Complex DNA mixture as a model system for information flow in biomolecular systems

1. Identification of ssDNA sequences representing accurate information that is embedded into various DNA mixtures to create a model system.
2. Singleplex and Multiplex PCR to isolate and specifically amplify a stretches of DNA-gene replicate that contain the intended information sequence.
3. Insert this gene replicate to a linearized dsDNA sequence (2344 bp) and circularise the

whole sequence to form a plasmid with gene replicate.

4. Transform plasmid with gene replicate into bacteria for expression and multiplication.

This expressed and multiplied plasmid represents a complex DNA mixture that is extracted and purified to form a model system for accurate information transfer in biomolecular system.

4.1.2. Identification of accurate information sequences

Specific ssDNA sequences that represent accurate information are identified and embedded into various DNA mixtures to create a model system.

To achieve first goal of this study, various ssDNA sequences of lengths 40-100 nt are identified from a lambda DNA genome. All these eight sequences represent the accurate or intended information, here after described as specific ssDNA sequences. These ssDNA sequences will be embedded into a complex DNA mixture of varying complexities.

- **Sequence 1: 25 nt length**

  TATAAGGGGATGTATGGCGATGTGG

- **Sequence 2: 30 nt length**

  TATAAATTCTGATTAGCCAGGTAACACAGT

- **Sequence 3: 35 nt length**

  AGATGTTGAGCAAACTTATCGCTTATCTGCTTCTC

- **Sequence 4: 40 nt length**

  GTTCGCGGCGGCATTCATCCATCCATTCGGTAACGCAGAT

- **Sequence 5: 50 nt length**

  CGGGTATCCAGCTTCTCCTTGACGGCTTTGAAGGAACGGAACAGC

- **Sequence 6: 55 nt length**

  ACGCCAGTCGCCACTGCCGGAGCCTTCATAAGCAATATCAACAACGACGG

- **Sequence 7: 60 nt length**

  CAAAATTGAAATCAAATAATGATTTTATTTTGACTGATAGTGACCTGTTCGTTGC

- **Sequence 8: 80 nt length**

  GGATGTTACGTCATAAAGCCATGATTCAGTGTGCCCGTCTGGCCTTCGGATTTGCTGGTATCTATGACAAGGATGAAGC

4.1.3. Singleplex and Multiplex PCR for specific amplification

Various PCR techniques are applied to isolate and specifically amplify stretches of DNA-genes that contain the specific ssDNA sequences.

Using lambda DNA genome as template, four dsDNA sequences which constitute specific ssDNA sequences are first separately amplified using Singleplex PCR and then collectively amplified using Multiplex PCR. Along with the intended PCR products, unspecific products were also present. All the PCR products are analysed with agarose gel electrophoresis, and dsDNA products are marked with Ethidium Bromide staining and visualised with UV illuminator. These dsDNA products are shown in Fig. 4.3 and 4.4. Fermentas O range ruler ladder 50 bp is used to identify and quantify the GR sequences with respect to length.

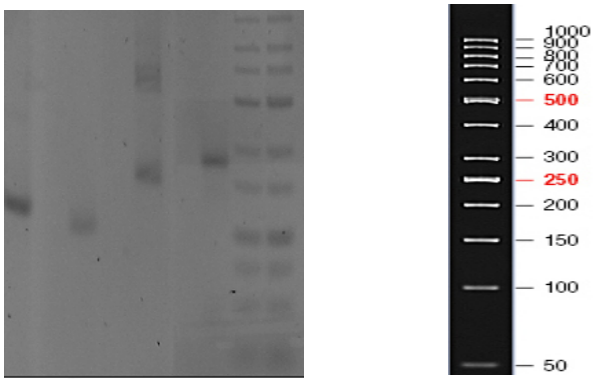
Specific PCR products based on length are excised from gel and purified with gel extraction kits from Jena Bioscience. They are run again in agarose gel and checked for successful gel extraction and purification. The concentration of PCR products is checked with Nanodrop spectrophotometer. Related protocol for agarose gel electrophoresis, gel extraction of dsDNA and purification can be found in Methods section.

Figure 4.1 shows four PCR products of various lengths run in an agarose electrophoresis. These sequences are purified from agarose gel to remove excess of primers and enzymes from PCR product. Using unpurified PCR products for downstream applications may decrease the stability of reaction.

Listed are lengths of various PCR products. Fermentas 50 bp gene ruler ladder is used to estimate and quantify the lengths and concentrations of PCR products.

Lane 1: product 1-107 bp **Lane 2**: product 2-128 bp

Lane 3: product 3-153 bp **Lane 4**: Product 4-177 bp



Lane 1 2 3 4 5 6

Figure 4.1: Singleplex PCR

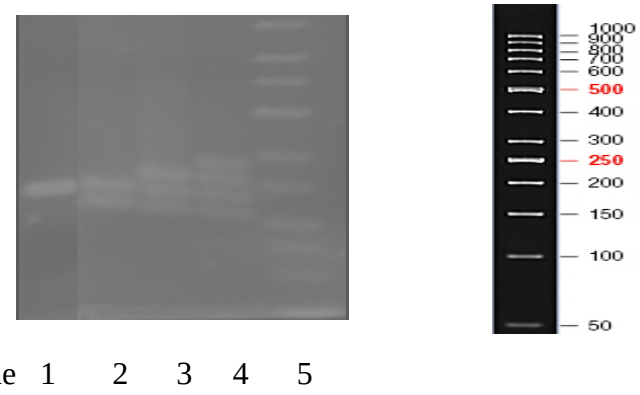


Figure 4.2: Multiplex PCR

Figure 4.2 shows an agarose gel electrophoresis picture of four multiplex PCR products.

Lane 1: Product 1

Lane 2: Product 1 + Product 2

Lane 3: Product 1 + Product 2 + Product 3

Lane 4: Product 1 + Product 2 + Product 3 + Product 4

Speciality of multiplex PCR is to amplify and isolate multiple target regions from multiple templates in a single reaction.

All these PCR products represent gene replicates. Our specific ssDNA sequences are part of these gene replicates. Next part of the project is to insert these gene replicates into a vector and use genetic recombination to add intrinsic and extrinsic noise sources to specific ssDNA sequences. This makes our model system more complex in nature and molecular dynamics unpredictable, as there are many regions in complex system, which exhibit various degrees of similarity to intended information sequences. This makes the retrieval procedure more complex.

4.1.4. Cloning of gene replicates

This gene replicate/dsDNA PCR product is cloned into a linearised dsDNA sequence (2974 bp) to form a plasmid vector.

To increase the complexity of DNA mixture, gene replicates are added to the multiple cloning site of a 2974 bp linearised vector. Vectors are generally Plasmids which can be transformed into a host, in our case E.Coli bacteria and successfully express and multiply it inside the host. This is genetic engineering/recombination. For this purpose Clone JET PCR cloning kit is used. Specific PCR product-3 (GR-3) is ligated with the multiple cloning site of the linearised vector sequence provided with the kit and subsequently a circularised plasmid is obtained. Multiple cloning sites represent an open ended region that binds a PCR product to itself and becomes circular.

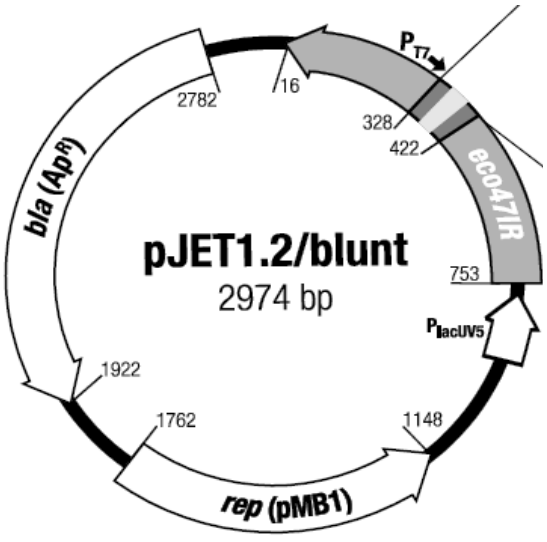


Figure 4.3: pJET1.2 vector map

Figure 4.3 shows a schematic of pJET1.2-blunt vector, which has a length of 2974 bp. This vector is already made linear by cutting at a particular site. Dark grey regions on either side

of vector represent recognition sites for PCR Products. Ligation is accomplished with DNA Ligase. This vector is suitable only for blunt end PCR products. If a PCR is performed with Taq polymerase, then dA-overhangs are added to both ends of amplified sequence. So a blunting steps to remove the additional dA-overhangs. When dsDNA-PCR products are successfully ligated to the multiple cloning sites, the linear vector becomes circular with Gene Replicates/dsDNA products. This circularised plasmid vector is ready for molecular cloning and transformation into E.Coli bacteria.

4.1.5. Bacterial transformation plasmid with gene replicates

The plasmid vector obtained from cloning containing dsDNA PCR products are transformed into bacterial cells for expression and multiplication.

For bacterial transformation, TOP10 competent cells from Invitrogen are used. Competent cells have an altered thin cell wall that allows foreign DNA to enter easily. E.Coli is most suitable for competent cells. Our circular plasmid is made to enter E.Coli cells using heat shock method. Then the cells are plated on Agar plates with suitable antibiotic, in our case Ampicillin. These Agar plates supply necessary nutrients for the cells to survive and multiply. They are stored at 37^0C overnight.

Related protocol for preparation of competent cells from E.Coli and heat shock protocol for bacterial transformation can be found in the Methods section in detail.

If transformation is successful, white colonies can been seen on plates. These colonies contain successfully transformed plasmid in E.Coli bacteria. These colonies are numbered and analysed with colony PCR. Five of the seven colonies were positive which means only these five contain the plasmid of interest. Now, the colonies are cultured and plasmid DNA is extracted with plasmid extraction kits. Figure 4.4 shows colonies on LB-Agar plates after overnight incubation in 37^0C incubators.

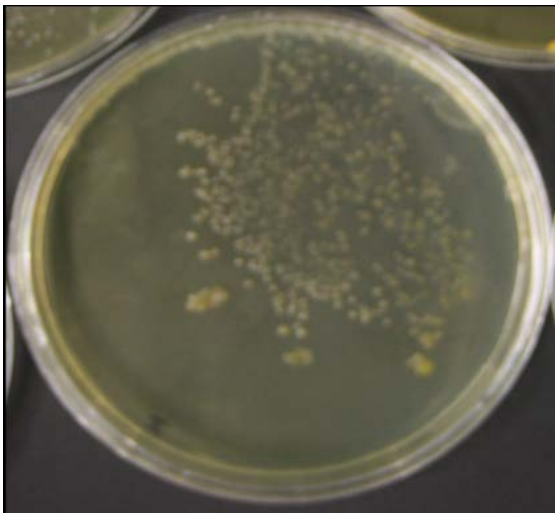


Figure 4.4: E.Coli colonies after successful transformation in a petri dish

These colonies are used for colony PCR as templates, as proof for successful transformation. For this colony PCR primers provided in the cloning kit are used. If colony PCR is successful, then a 250 bp product should be amplified by PCR. This 250 bp product contains the intended dsDNA PCR sequences/gene replicates. Figure 4.5 shows that transformation is successful. Lanes 2-6 show colony PCR products. Now the colonies are cultured in LB medium with Ampicillin as antibiotic. The cell are grown and extracted with standard bacterial culture procedures. Information about cloning kit and protocol can be obtained in Methods section.

Figure 4.5 shows agarose gel electrophoresis of complete model system.

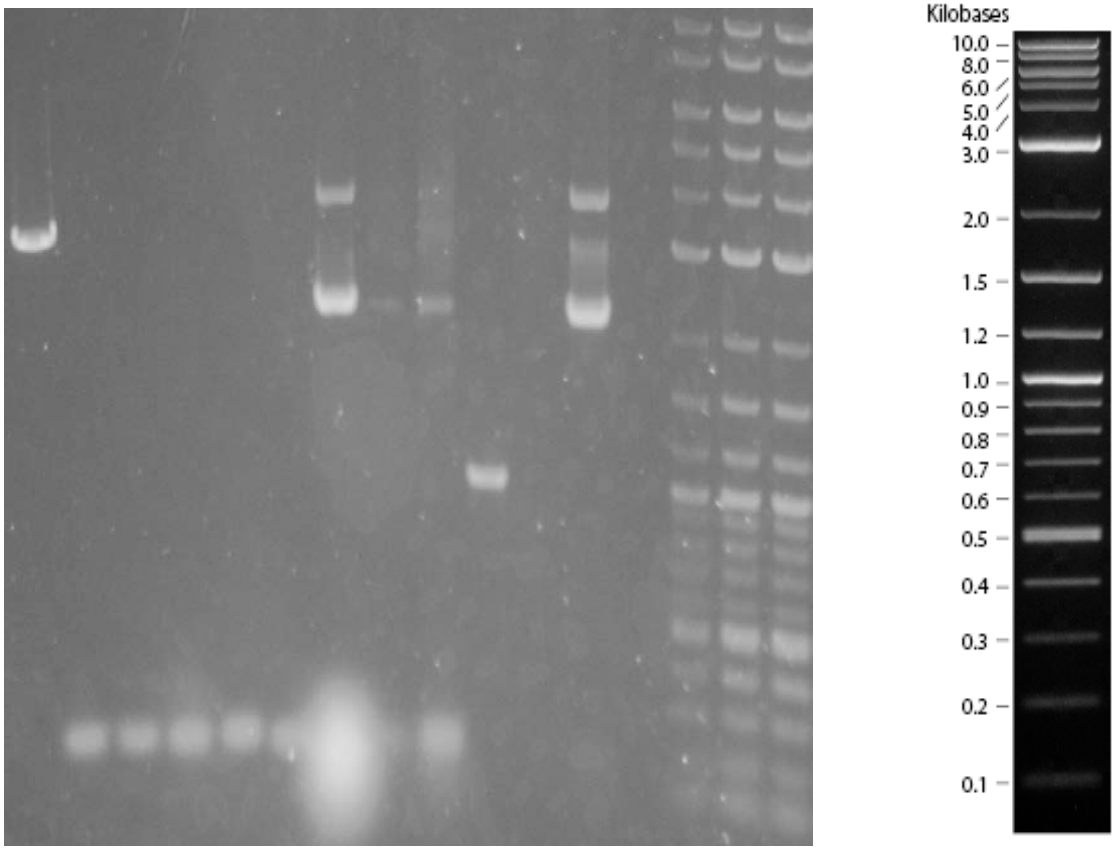


Figure 4.5: Agarose gel representation of a model system- proof of successful plasmid transformation

Lane 1: Linearized plasmid.

Lane 2-6: Gene replicate/dsDNA PCR product of length 153 bp.

Lane 7: Intended plasmid with RNA artefacts as a smear at the bottom of the lane.

Lane 8-9: PCR products for gene replicates with plasmid as template after extraction and purification. As plasmid generally has very high concentration, it was also visualized in the gel.

Lane 10: Control PCR product 944 bp.

Lane 12: Intended Plasmid-Complex DNA Mixture after extraction and Rnase treatment.

Lane 14-16: 1 kb ladder from New England Biolabs is used as reference.

From Figure 4.5 it is clear that the linear plasmid is generally seen as a single band in the

gel. When circularized, part of the plasmid often assumes a super-coiled structure. So there are two bands. Super-coiled plasmid is smaller in size than a circular plasmid, so it can be seen below the circular plasmid in agarose gel.

Now the circular plasmid represents our model complex biomolecular system. Extrinsic and intrinsic noise sources are added due to various processes like replication and division inside the cell. They have an influence on gene expression levels.

4.2. Representation of probable noise sources in complex DNA mixture

For successful retrieval of specific information from complex DNA mixture, accurate hybridization of an oligonucleotide sequence (17-20 nt) to its complementary region is necessary. An algorithm was developed to show the probable noise sources in the complex DNA mixture. Figure 4.6 shows various probabilities of unspecific binding. Binding score represents the probability that the primer can bind to any 20 nt sequence in the complex DNA mixture. 100% binding score represents the binding of primer to its specific complementary sequence. Figure 4.7 shows the binding probabilities between 55-100%.

It can be understood that the regions before and after the primer binding site are more probable sites for unspecific hybridization. This means that primer binding to any of these unspecific binding sites results in a change in length and nucleotide composition of the resulting sequence when this sequence data is analysed with hybridization based information retrieval techniques, false information is retrieved. The software part for this algorithm was developed by Srikanth Duddela, Helmholtz centre for infection research.

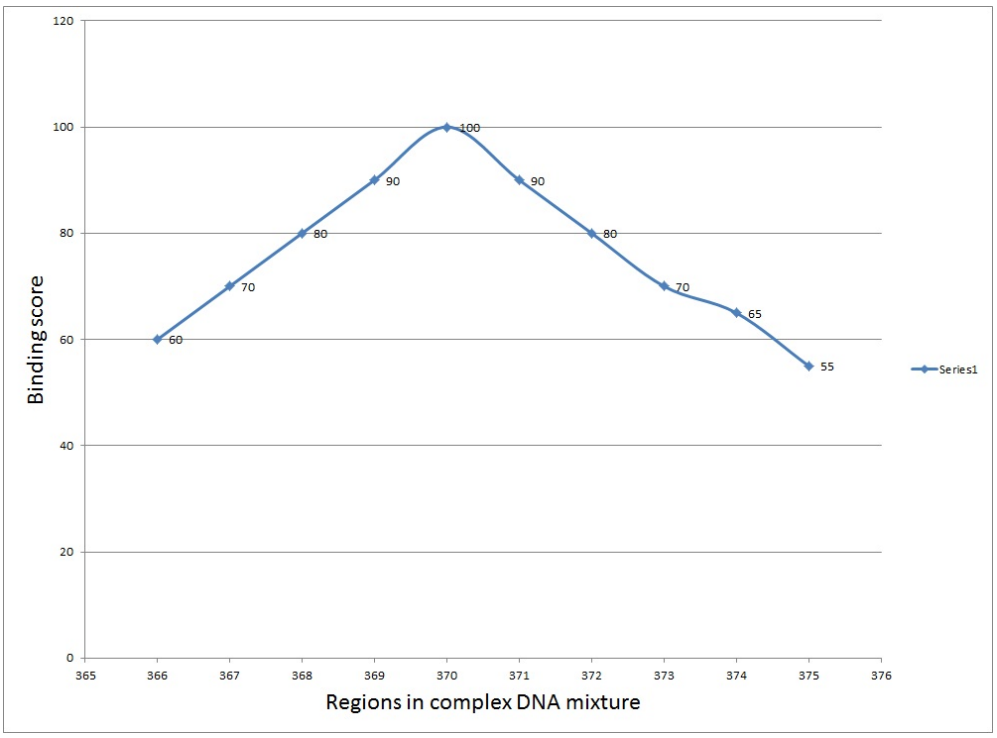
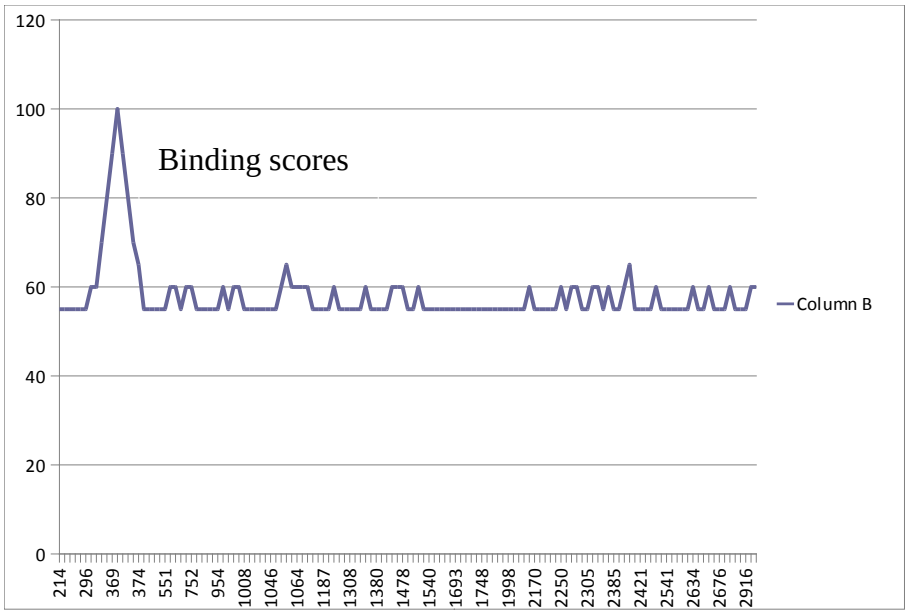


Figure 4.6: Probabilities of unspecific binding



Regions in Complex DNA Mixture

Figure 4.7: binding probabilities between 55-100%

4.3. Retrieval of specific ssDNA sequences from complex DNA mixture

Second part of our work involves accurate identification, retrieval and analysis of multiple specific ssDNA sequences from successfully realized complex DNA mixture or model system. The location of specific ssDNA sequences in the model system is known through DNA sequence analysis.

4.3.1. Identify specific ssDNA sequences directly from complex DNA mixture

Various existing techniques were tried. Prominent among them are Linear PCR and isothermal amplification, linear PCR is performed with varying number of primers, targeting multiple ssDNA sequences in the circular plasmid. The linear PCR products were run in PAA-Urea 12% gels and silver stained to identify ssDNA sequences.

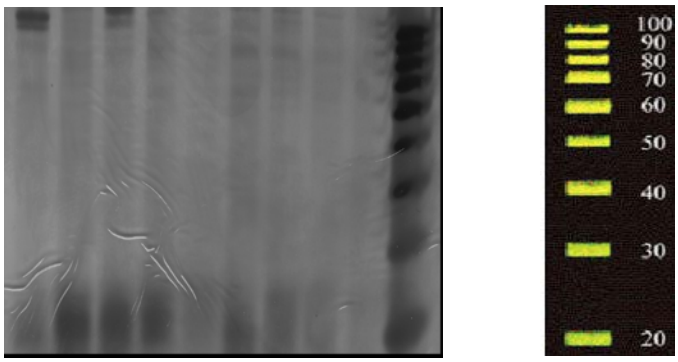


Figure 4.8: Unsuccessful Linear PCR experiments-no clear products

Fig 4.8 shows linear PCR experiments with a single primer for a sample lambda DNA template. It can be clearly seen that there are multiple bands of weak intensities below 100 nt range. In ideal case, there should be only one band/product of a specific length and no other products. But presence of weak/zero bands shows the inherent weakness in linear PCR experiments. So it can be concluded that stopping Taq polymerase activity depending on length distribution is difficult and reaction dynamics become unpredictable.

❖ **Linear PCR experiments**

The linear PCR products are shown in Fig. 4.9. Here random primers are used for shorter template of length below 200 bp. This shows a random distribution of multiple products in the gel matrix. This is approximate length of a gene. It can be clearly seen that there are products below 100 nt. But, the exact sequence is not known and fidelity and reproducibility of these experiments cannot be predicted accurately. Intention of this experiment is to identify only one band in the gel matrix.

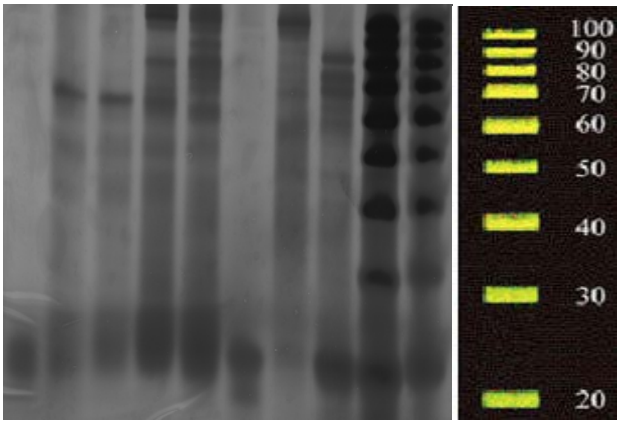


Figure 4.9: Linear PCR experiments with random primers

❖ **LATE PCR experiments**

With LATE PCR, two primers were used. The primer sequences are specified in the Protocols section. The primers are mentioned in the protocols section. One primer has limited concentration of 0.01 pmol/uL, while the other primer has a concentration of 100 pmol/uL. So initially, there is exponential amplification. But as the limiting primer is used up, there is only one primer for amplification and this result in linear amplification.

The results are shown in Figure 4.10. Initially there is only one band that signifies exponential amplification (highlighted as *). This refers to specific product. Then there is a second band below the first band (**). This is a linear PCR product. But it is not possible to limit the linear PCR product length to below 100 nt as the polymerase extension step cannot be stopped abruptly. New alternatives, like abrupt temperature increase/cool-down were tried. However, the results were not reproducible and sometimes ambiguous.

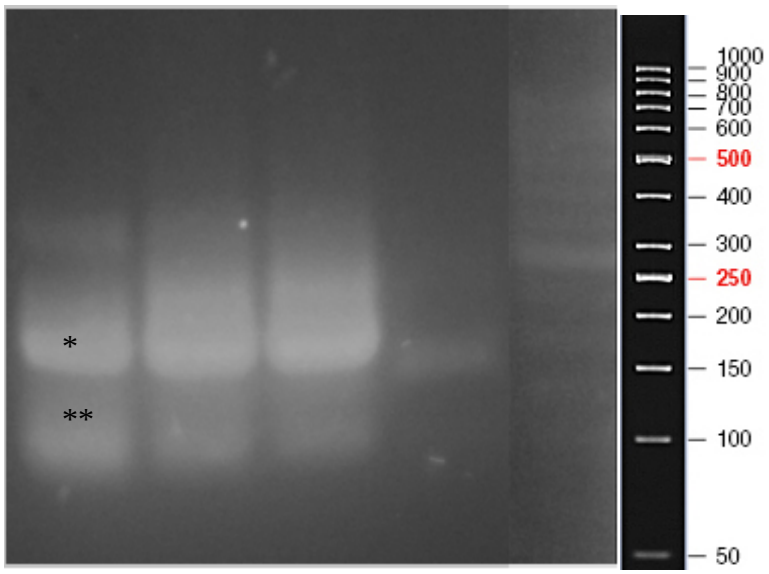


Figure 4.10: Late PCR-for linear amplification after exponential amplification

4.3.2. Conclusions from linear PCR and asymmetric PCR experiments

It is not advisable to limit the reaction based on shorter time frames

This means that, in order to obtain specific ssDNA sequences, which is essentially a 40nt

long ssDNA sequence; the extension time of linear PCR should be 3-6 sec. commercially available enzymes are optimized for longer amplified sequences in a shorter time frame. For example, Taq polymerase has an extension of 1 Kilo-base pair per one minute. So reduction of extension time frame to 2-5 seconds resulted in failure of the PCR reaction and subsequent false results. These results are shown in Figure 4.11.

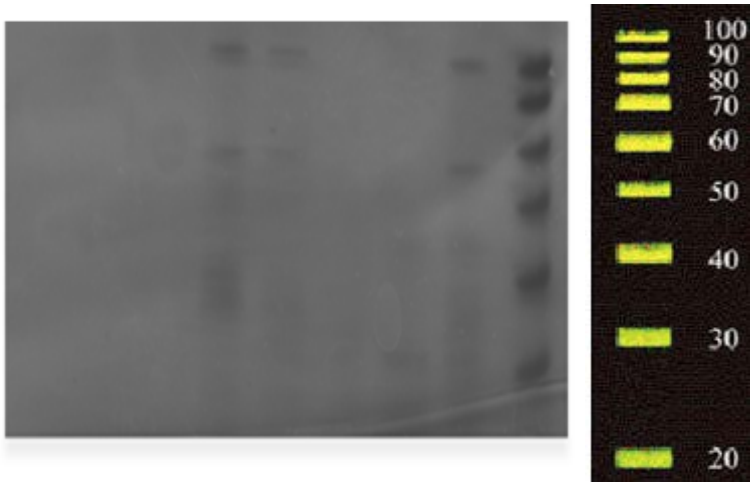


Figure 4.11 shows Linear PCR experiments with very short extension times

- **Conclusion**

There is an inherent shortcoming in ssDNA amplification experiments with PCR. Due to this, mispriming or identification and amplification of wrong targets are highlighted. By using two primers, in conventional PCR there is always amplification of specific product from the template along with unintended products which can be easily identified and separated. As only one primer is used in linear PCR, there is always mispriming or unspecific binding on primer to template. Thus, the fidelity of our experiments is adversely affected.

General PCR methodologies depend on a standard PCR protocol. It includes an annealing step at 50-60^{0}C and then an extension step of 72^{0}C. Annealing step means binding of 17-25 oligonucleotides to the denatured ssDNA template. Then, during extension step dNTPs bind to the template and from a complementary strand.

❖ **Modifications tried with linear PCR experiments**

In conventional PCR-Thermal cyclers, used for our purpose there is an inherent problem with ramping. The sample is always in contact with the heating block, during the time lag between various Annealing and extension steps. So there is always extension of complementary strand bound to the template. This results in ssDNA sequences that are longer than expected.

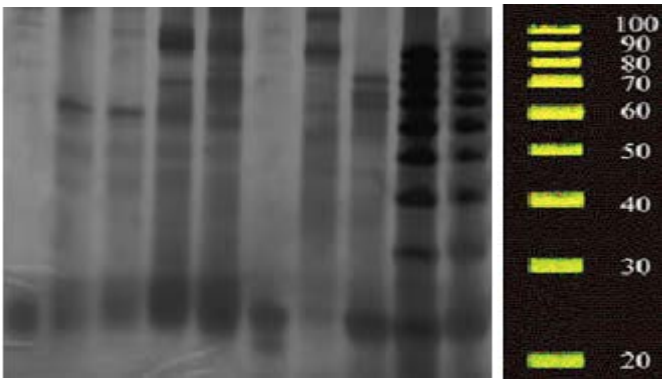


Figure 4.12: Linear PCR experiments with temperature ramping

Our initial efforts concentrated on including an extra temperature step of 40C between annealing and extension steps. This gives the possibility for the heating block to cool quickly and stop the extension of complementary strand. They were successful, as the

lengths of intended sequences decreased. These results are shown in Figure 4.11. A linear PCR experiment with temperature ramping was performed eight times for the same template with same primer. All the eight lanes show products at different lengths. Sometimes, there were multiple products from the same reaction.

- **Conclusions**

It can be concluded that various new approaches to identify ssDNA sequences and successfully amplify them were tried out and were successful to an extent. **However, most of the findings were inconclusive**. The extension ability of the enzyme was not hindered completely with the addition of the temperature ramping step. This renders the reaction dynamics extremely unstable. Thus, the reaction can be affected due to various factors like lack of reliability, fidelity and repeatability. This shows a need for development of a methodology that involves a stepwise procedure to obtain successful, repeatable and conclusive results.

4.3.3. Initial experiments with FPLC for separation of ssDNA sequences

After multiple linear PCR experiments were performed, FPLC analysis was used to separate and identify linear PCR products with multiple columns. Initially a Waters oligo separation column-XTerra with C_{18} material was used. The protocols and methods used to separate our PCR products were described in the Methods section. Our laboratory has an Amersham GE Akta 900 FPLC which reaches 5MPa of pressure and flow rates upto 20 ml/min.

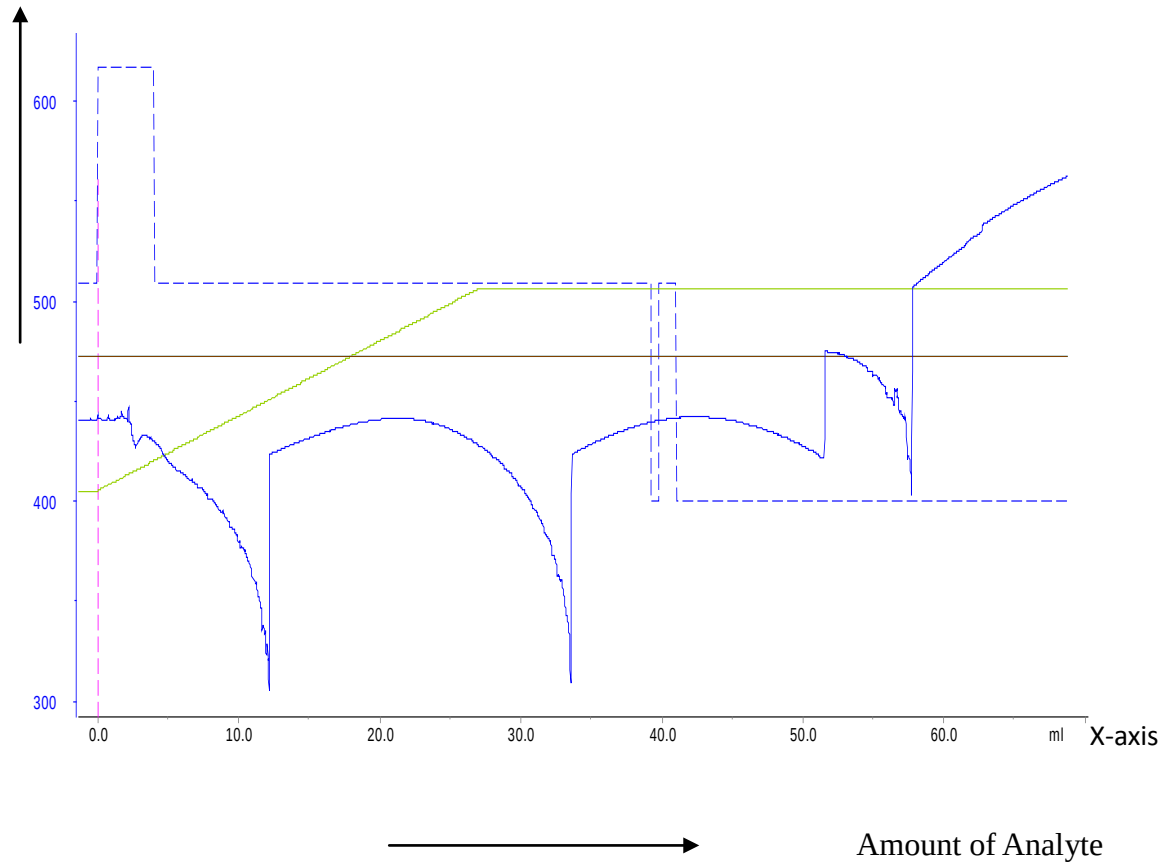


Figure 4.13: Chromatogram for separation of multiple ssDNA sequences in a solution

Our initial efforts were to modify this FPLC for separation of ssDNA. Figure 4.13 shows chromatogram for separation of ssDNA of various lengths between 20 to 100 nt. These chromatograms show various peaks which are not conclusive. Ideally, there should be a multiple clear peaks, but this is not the case here.

Reasons for inconclusive results are presented here

1. Single-stranded biomolecules have an inherent tendency to form loops. This may generally result in many problems such as shorter retention times, unpredictable binding of looped biomolecules to the materials in column.

2. The solution to this problem is to heat the column to a temperature above the usual melting temperature of ssDNA. This results in linearization of ssDNA and subsequent correct analysis. A self-designed oven was fabricated in our university mechanical workshop. This is cost effective and it is possible to use various temperatures in between 10 to 80 °C. However, this effort was not successful. Then various other modifications were tried.

3. Any biomolecules that are needed to be separated were mixed with analyte, in our case 10% Acetonitrile and injected into the column. To push the biomolecules through column, all chromatography equipment has pump to generate the necessary pressure. Flowcharts for separation of DNA, RNA, peptides and proteins are different to each other. Various flow rates were used to make separation of ssDNA products possible. But this was successful only to extent due to technical constraints. HPLC, which generates higher pressure, would be more suitable for our experiments. This may have resulted in initial wrong analysis.

5. Need for new strategy

The above conclusions show the need for a new strategy for accurate identification of specific ssDNA sequences. This involves the following steps.

Stepwise model for accurate information retrieval

1. Identification of specific ssDNA sequence location in complex DNA mixture, which is essentially a circular plasmid expressed in E.Coli cells.
2. Isolating larger dsDNA region which is a PCR product that encompasses specific ssDNA sequences.
3. Identification and retrieval of multiple specific ssDNA sequences from larger dsDNA region through our MEA technique. This contains various noise and unspecific signal sources.
4. Checking for possible presence of specific ssDNA sequences with HPLC.
5. Isolation of specific ssDNA sequences based on length distribution with PAA-Urea gel electrophoresis
6. Retrieval of specific ssDNA sequences through various molecular biology techniques and our modified biotin-based nucleic acid extraction technique.
7. DNA microarray experiments.
8. Successful verification of specific ssDNA sequences through membrane based hybridization methods.
9. Proof of successful retrieval of specific *ssDNA sequences-accurate information* through Sanger sequencing.

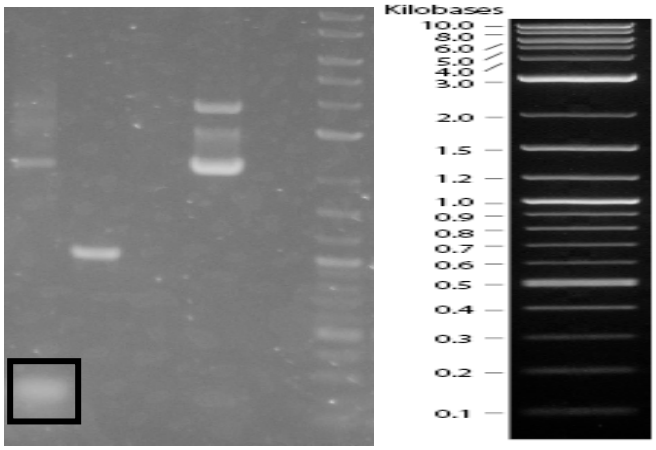
6. Stepwise model for accurate information retrieval

6.1. Identification of specific ssDNA sequence location

From the complex DNA mixture, prepared in Project 1, the location of specific ssDNA sequence-accurate information is inserted by molecular cloning into the plasmid.

6.2. Isolating larger dsDNA region

So the conventional Singleplex PCR is used to identify the larger region, which encompasses the specific information sequence. For this, two primers are designed with Primer3 software to identify, isolate and amplify larger dsDNA region specifically from the complex DNA mixture. The PCR product is subjected to agarose gel electrophoresis and amplified larger dsDNA region is identified and extracted from the gel. Four larger dsDNA regions are used in this dissertation.



Lane 1 2 3 4 5 6

Figure 6.1: Agarose gel electrophoresis-shows the larger dsDNA region from complex DNA mixture

Lane 1: linearized plasmid and amplified larger dsDNA region marked in black region which encompasses the multiple cloning site-specific information sequences.

Lane 2: shows control PCR product 944 bp.

Lane 4: shows Complex DNA mixture-Plasmid

Lane 6: 1 kb ladder

6.3. Application of MEA technique

Identification and retrieval of multiple specific ssDNA sequences is accomplished through our MEA technique. Our novel modified end amplification method was explained in the chapter 3.3 of the Methods section.

6.4. Verification for possible presence of ssDNA sequences

Reverse Phase High Pressure Liquid Chromatography was used to separate the unpurified linear PCR product. The results are shown in a chromatogram. It shows the concentration of various biomolecules present in the solution. In this case, ssDNA sequences are eluted through a column. Waters Xbridge OST C18 column is used for this purpose.

Biomolecules bind the column and when polarity is decreased, the biomolecules are separated from the column and are eluted. Retention of the biomolecules to the column is dependent on polarity of biomolecules.

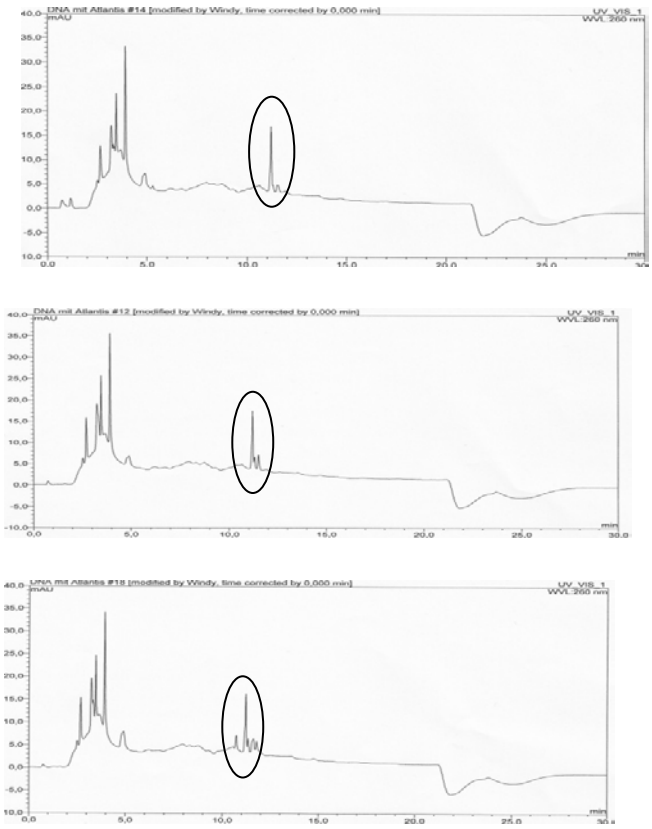


Figure 6.2: Chromatogram of multiple ssDNA sequences separated from complex DNA mixture

Figure 6.2 shows various chromatograms for ssDNA separation with multiple peaks. In the reaction mixture obtained from MEA technique explained in the Chapter 3.3 of Methods section. There are various clear/longer peaks and several shorter repeating peaks. Peaks which are of interest and indicate the possible presence of ssDNA are marked in black. This shows clear evidence of presence of ssDNA. But, to estimate the length of the ssDNA sequences and judge the fidelity of obtained sequences, these data is not conclusive.

6.5. Retrieval of specific ssDNA sequences through our modified nucleic acid extraction technique

To further identify and isolate specific ssDNA sequences from larger dsDNA a region, further screening is done based on PAA-Urea gel electrophoresis. Here there is conclusive evidence, to support our claim that intended specific ssDNA sequences were accurately

identified and amplified from dsDNA regions.

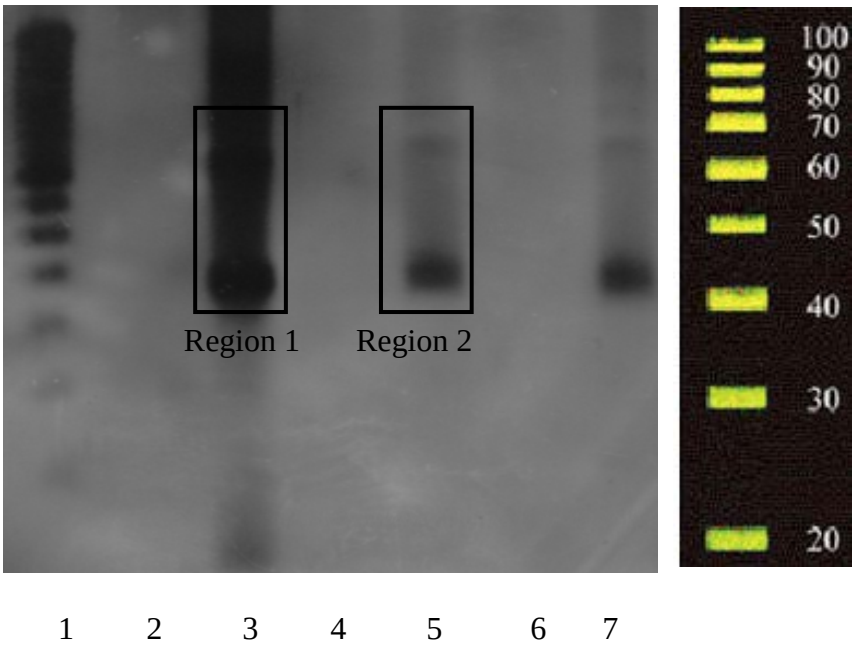


Figure 6.3: MEA technique products before and after purification

Figure 6.3 shows PAA-Urea 12% gel after silver staining. Lane 1 shows a 10 bp O range ruler ladder. Lane 2 is unpurified MEA product. Lanes 3 and 5 are purified modified PCR products with biotin Streptavidin extraction and zymo ssDNA columns. It can be clearly seen that after purification, there is no smear in gel bands. This smear is from some unspecific products and also due to fragmentation of template from high urea concentration. After purification, most of the impurities are eliminated and the purified product is suitable for downstream applications.

Figure 6.4 shows plot depicting noise regions between unpurified and purified regions in Figure 6.5. The regions are depicted in region 1 and region 2.

Region 1: smear/noise

Region 2: after purification

It can be clearly seen that peak marked in Region 1 is much higher than peak marked in Region 2. Region 1 corresponds to noise in unpurified product and Region 2 corresponds to reduced noise and mispriming or other false hybridization factors. From a preparative gel analysis and commercial columns it is difficult to remove unspecific products or noise completely.

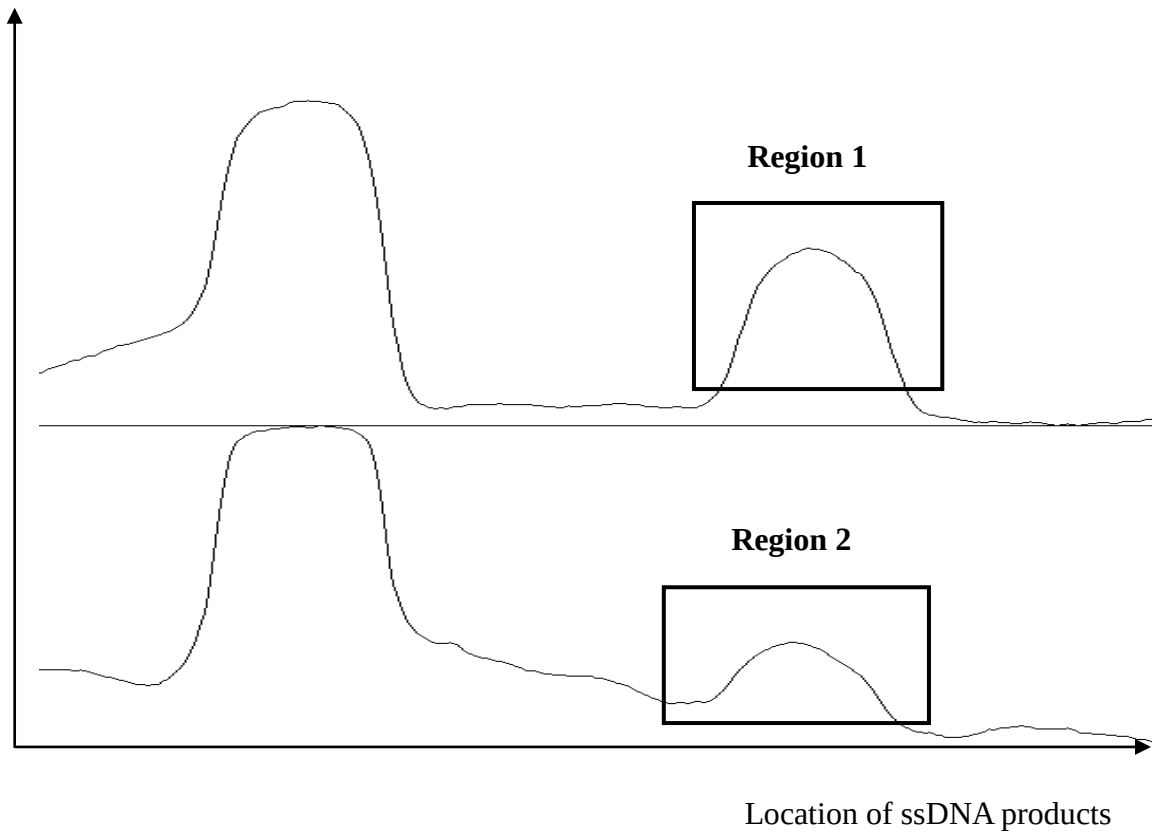


Figure 6.4: Difference between unpurified and purified regions in MEA products-curves plotted with band intensity (y-axis) to (x-axis) location of ssDNA products in the gel lanes.

These results show that our MEA method is very successful in identifying, isolating and amplifying short ssDNA sequences of specific length and predefined nucleotide composition. But, high throughput applications can analyse multiple ssDNA sequences at the same instant. Therefore, for further viability, our method is extended to identify multiple ssDNA sequences from different regions of the same template.

To further extend the application of our novel MEA method, two primers were used to identify two regions in a same template. This is used to identify, isolate and amplify multiple specific ssDNA sequences from the same template.

Primers are designed in a way that they bind at different regions of the template and extend till the end of the template. Figure 6.5 shows the gel electrophoresis separation of the multiple ssDNA sequences. One of the ssDNA sequences has a length of 40 nt and other sequence has a length of 80 nt.

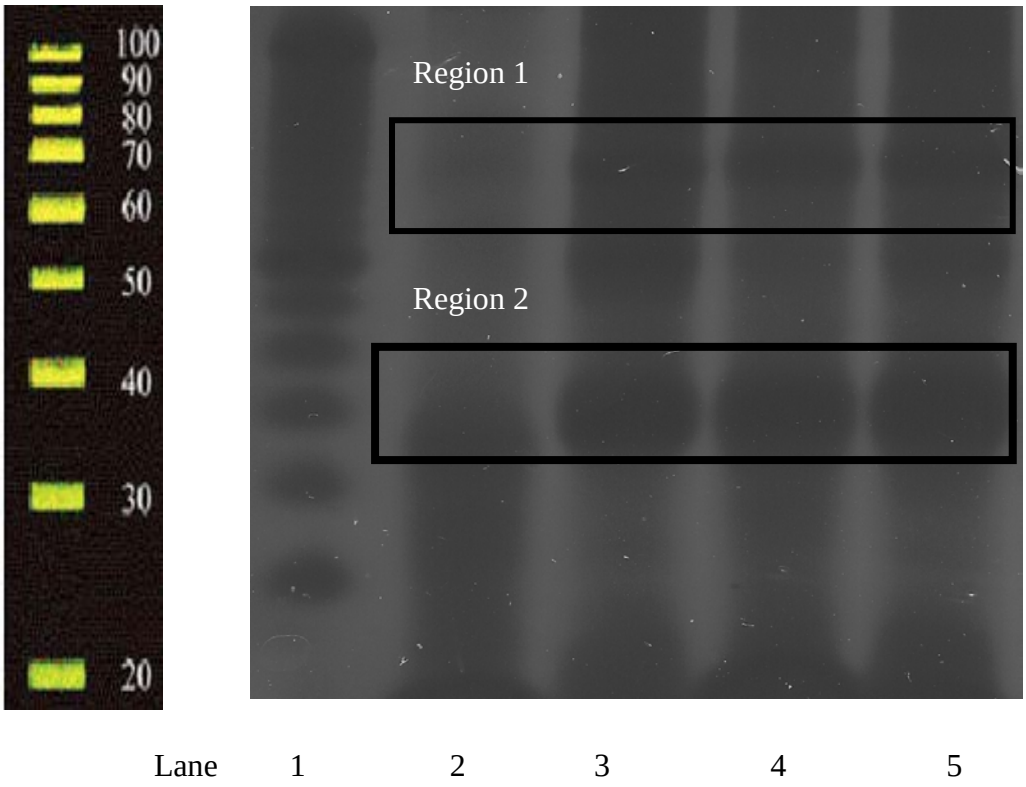


Figure 6.5: Multiple specific ssDNA sequences from a single L.R. 40-80 nt

40 nt products are identified as region 1 and 80 nt products as region 2 in the gel.

It can be clearly seen that 40nt sequences are identified from multiple dsDNA regions in complex DNA mixture. In Figure 6.6, lanes 1-4 show multiple ssDNA sequences from

multiple LR3 regions which are marked as region 1. For additional fidelity verification, 40 nt sequence was purchased from Metabion GmbH and also run in the gel. It is shown in lane 6 and marked as region 2. It can be seen that the intended 40nt sequence and pre-synthesized 40 nt sequence are at the same level in the gel.

This presynthesised sequence has the same sequence composition as intended 40 nt sequence and has the same molecular weight. So ideally, in a gel they should be at the same level. This further proves the accuracy of MEA experiments.

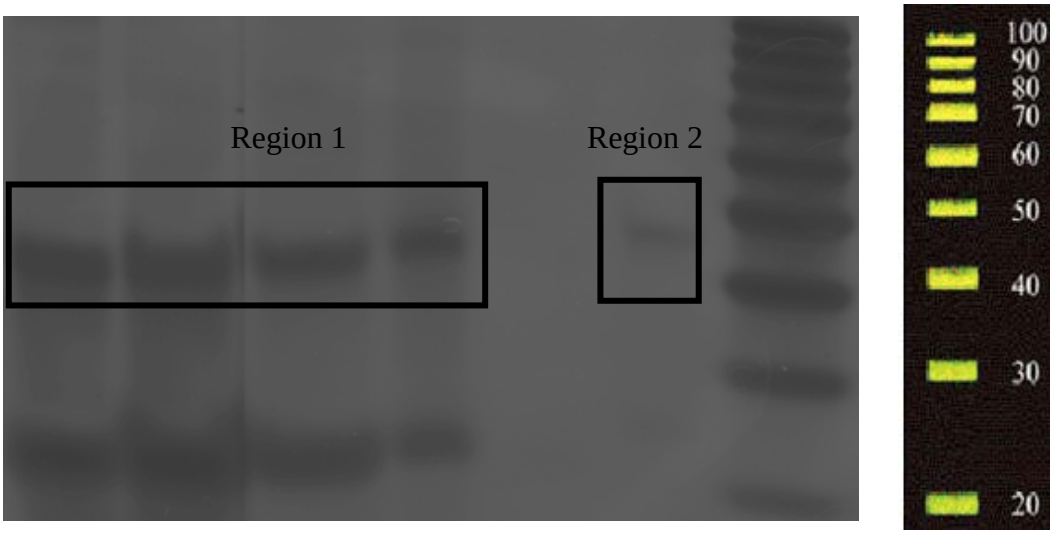


Figure 6.6: 40 nt/specific ssDNA sequences based on length. Identified and amplified from LR3

This method can be used to identify, isolate and retrieve ssDNA sequences of any length and can be used for difficult templates with high GC content. To vindicate this argument, ssDNA sequences of various lengths were amplified with MEA method and separated with PAA-Urea 12% gel. Figure 6.7 shows multiple ssDNA sequences of various lengths from 25 nt to 60 nt. It can be seen that almost all the sequences are free of noise and mispriming. Lane 3 shows an unpurified PCR product to further show the difference between purified and

unpurified products.

It can be concluded that identification and retrieval of specific sequences from a complex DNA mixture significantly reduces various errors and false positives that can occur due to random fragmentation, sample handling and data verification.

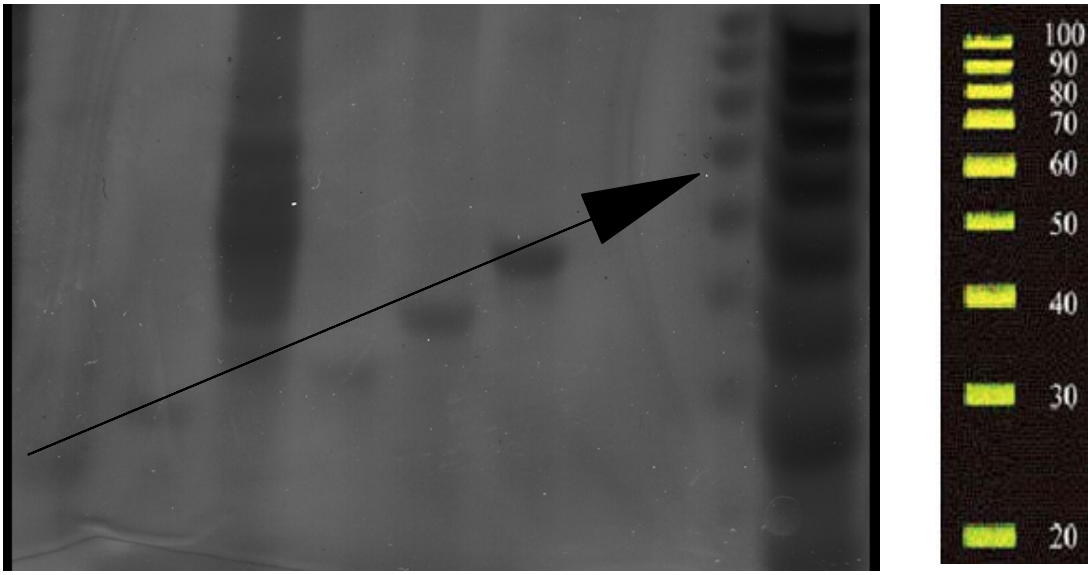


Figure 6.7: Multiple specific ssDNA sequences of various lengths from 25 nt to 60 nt from complex DNA mixture.

6.6. Microarray experiments

DNA microarrays are conventionally synthesized in 3´-5´ direction. Also, the oligonucleotides have a 5´-end NPPOC modification which can be cleaved upon illumination. This allows for predefined binding of nucleotides. This is conventional and correct synthesis procedure.

For the following experiments, the direction of the sequence for synthesis on to the microarray surface is accidentally provided in wrong/opposite way due to a programming error. The sequence for immobilization is provided in 5´-3´ direction instead of 3´-5´ direction.

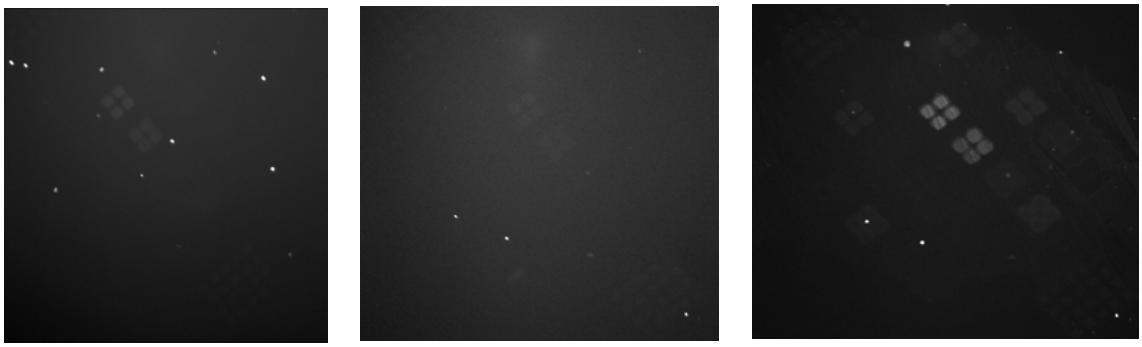


Figure 6.8: Multiple microarray fluorescence images

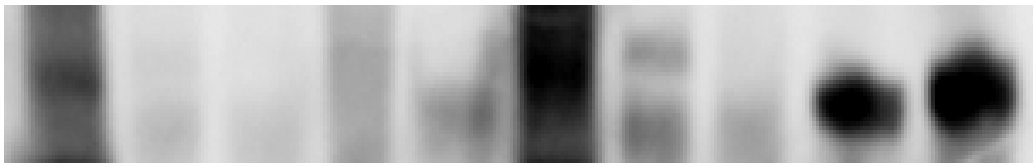
However, the microarray pictures from Figure 6.8 show the presence of hybridization and fluorescence signals which are consistent. In all the three microarray pictures, there are hybridization intensity signals from the same region of microarray. Here three targets prepared from different ssDNA synthesis techniques were used for the same microarray. This is a case of reverse binding or false hybridization. Also, other regions of microarray also show traces of hybridization. This is similar to previous research that relates to parallel hybridization. Similar experiments were performed by Heike Wech and Bjorn Ackermann from Department of Biological experimental Physics, Saarland University.

As it known that the design of immobilized ssDNA sequences on the microarray was not in accordance to correct procedure, the data retrieved from this microarray is difficult to interpret. The fidelity of the ssDNA targets produced from our modified MEA technique cannot be estimated from this data analysis. The array synthesis and hybridization parts of

the experiments were performed by Christian Trapp from Department of Biological Experimental Physics, Saarland University.

6.7. Successful verification of specific ssDNA sequences through membrane based hybridization methods

To check for accurate hybridization and thus accurate retrieval of information, Southern blot technique explained in Chapter 3.10 in the Methods section is used. This is one of the basic hybridization verification procedures especially for ssDNA. Our experiments used a modified version, where both targets and probes are ssDNA sequences.



Lane 1 2 3 4 5 6 7 8 9 10

Figure 6.9: Southern blot for 40 nt sequences

In Figure 6.9, Lanes 1, 2 and 3 show unpurified MEA 40 nt product in three different amounts (20, 10, 5 uL), lane 4 and 5 show ssDNA product from modified isothermal amplification method with single primer. Lane's 6-8 show purified MEA products. Lanes 9-10 show reference ssDNA sequences of 40 nt length and of same sequence as intended 40 nt sequence.

The Figure 6.9 shows hybridization signals from specific ssDNA sequences from blotting

experiments is all lanes. This proves that in membrane-based hybridization is perfectly suited for our experiments and purification of ssDNA. This is the simplest method and it eliminates many problems that are associated with microarrays and other surface based immobilization techniques.

Membrane-based hybridization techniques have the following advantages over hybridization methods using modified surfaces for ssDNA immobilization.

1. As the ssDNA sequences are printed on the membrane and get stuck between pores of membrane, problems with direction-based immobilization are completely eliminated.
2. Membranes can be heated to higher temperatures than glass surfaces, so longer targets can be used. With surface-based hybridization, higher temperatures can result in separation of probes from surfaces.
3. Manual printing completely eliminates any discrepancies related to wrong synthesis of oligonucleotides to surfaces.
4. Hybridization on membranes can be identified with many methods like Ethidium Bromide staining, fluorescent and non-fluorescent imaging.

Unpurified reaction products can be used for membrane based hybridization techniques. But importance of purification procedures cannot be completely neglected as removal of inhibitors like EDTA increases the specificity of hybridization.

6.8. Verification of signals with Sanger sequencing

The various ssDNA, which were successfully extracted and purified from complex DNA mixture, are sent for sequencing to find the exact sequence information. However, there

were lot of shortcomings in existing DNA sequencing techniques

1. It is suitable preferably for dsDNA sequences.
2. For ssDNA sequencing, the primers to be chosen for sequencing should be very specific and reduce mispriming and binding at different regions.
3. Most important shortcoming and especially important one in our case is that ssDNA sequences shorter than 100 nt cannot be sequenced.

So to convert ssDNA to dsDNA and also to test the accuracy of nucleotide composition with another hybridization technique, hybridization in solution is performed. This means, ssDNA sequence-target is hybridized with its complementary ssDNA sequence-probe in a solution based on standard hybridization protocol. Then the dsDNA is run on agarose gel electrophoresis and extracted and purified. Then this purified sequence is cloned into linearized vector and transformed into E.Coli cells. Figure 6.10 shows colonies on LB-Agar plates with Ampicillin.

Colony PCR was performed with Fermentas CloneJET kit with sequencing primers that amplify the region containing cloned hybridization product.

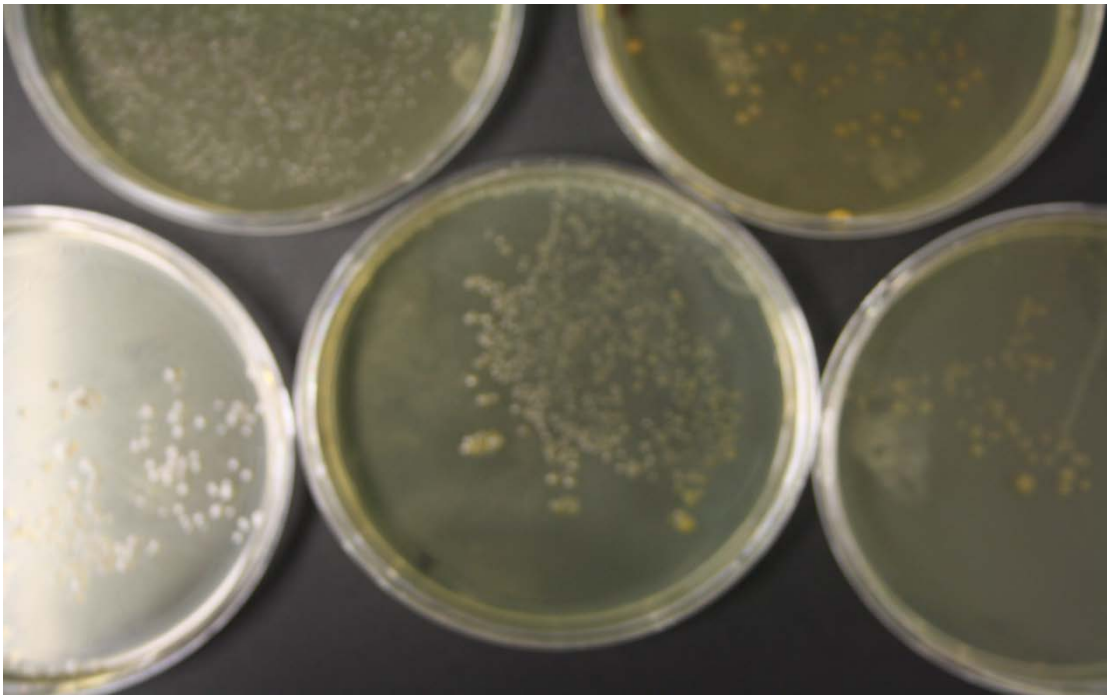


Figure 6.10: Colonies on LB-Agar plates with Ampicillin in a petridish

These 135 bp sequence has the required 40 bp sequence and is shown in a box Figure 6.11.

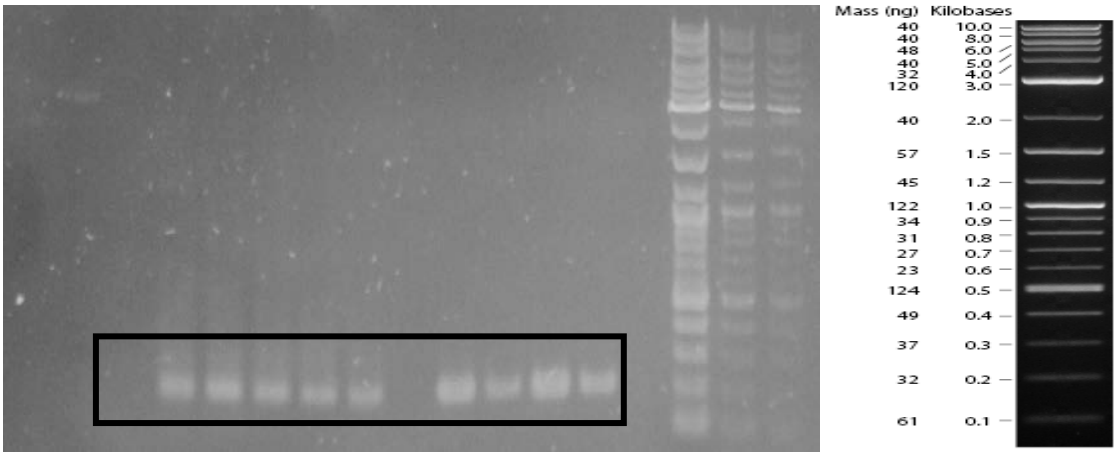


Figure 6.11: Colony PCR for 135 bp sequences that encompass the intended ssDNA sequences.

These 135 bp sequences are extracted and purified from agarose gel using various commercial kits. These sequences are sent to Gatc Biotech for Sanger sequencing at a

concentration of 15 ng/uL.

The results from sequencing are shown below. It shows the sequence of 135 bp sequences.

>sequence3-GATC-forwardprimer-710409 10256658

NNNNNNNNNNNNNNNNNNTCGANTNNNNNNANAT**GTTCGCGGNGGNATTCAT**

CCATCCATTCGGTAACGCAGATAGGATGAATGCGGTCCATTCGGTAACGCAGAT

AGGATGAATGCCGCCGCGAACATCTTTCTAGAAGATCTCCTACAATATTCTCAGCT

GCCATGGAAAATCGATGTTCTTNNNTCANTGTCTTTAATGCTGGNTCCGTTGCCN

CCCTCTTTATCGGGNCTTCCCACCTGCCCNGATGTCCTGCCTCGCCGGATTTCTGN

CCCCGCTTGCGCTTCGCGCTCTGCTTGCNAGCTTAGCAAANGAATCTCGGCGGAT

TCCNTNATCACNNNTNACNCGNNATGTCTGGGGNTTCAAANTCCCTCNACCTGA

ACNCNACTTCAACGCTGCNACCTTTCCTCCCTGCGCNNANAGGANAGTGAAGGC

TCGAGTGTGCGAGTGAACGGNCNNTTGCCCTCCNNNGAAATGATAGCGCCTACC

GGCNCCATCATCTGCTTCNCGATCTTACGCCNCTGTACCGTGGACTAAGACCGTC

CCAGCTNCTCTCTACNGTAGGGTGNTGTNCATANNNNTAGATACCCANAGGATNN

CTATCTGTGGTCNACCTGGACTTNCC

The highlighted region represents the required ssDNA sequences with predefined nucleotide composition. The sequenced product contains the intended information sequence that was embedded in complex DNA mixture. This clearly proves the fidelity and specificity of our modified MEA method.

Summary of results:

The realization of complex DNA mixture is shown in Section 4.1 (Results). Initially, specific information is identified from an information source and various stages of increasing complexities are added. Identification, retrieval and validation of specific information from a complex DNA mixture is based on a novel experimental approach. The effects of various constraints on information transfer are investigated. Constraints like sources for unspecific, cross hybridization play an important role in increasing the amount of raw data generated from conventional techniques.

The binding probabilities of possible noise sources are shown in Section 4.2. This shows that reducing unspecific binding is always important for specific information transfer.

The LATE PCR experiments which were carried out on a lambda DNA template are shown in Section 4.3. It can be clearly seen that the ssDNA sequences that are produced have lengths almost similar to the original template. The length of amplified sequences is much higher than optimum for oligonucleotide microarrays.

We concentrate on synthesis of ssDNA sequences of lengths between 20-100 nt. Also, these ssDNA sequences are validated with denaturing gels which keep the ssDNA linear through the presence of urea (denaturing agent). Urea breaks the secondary structures and loops in ssDNA. The stepwise model for accurate retrieval short ssDNA sequences synthesized with our MEA method is shown in Chapter 6. Section 6.5 shows ssDNA sequences of multiple lengths separated from a complex DNA mixture. This shows that this technique can produce specific data for various regions of a complex DNA mixture. So, various regions in the model system can be analysed in a single instant and the data is

converted to specific information through blotting and sequencing techniques. Figure 6.7 shows the length distribution of ssDNA sequences from 20-80 nt. This proves that our MEA method can synthesize ssDNA sequences of specific length as opposed to random fragmentation techniques.

In Section 6.6, parallel binding of DNA sequences is shown. This is against the inherent property of DNA-DNA binding, where ssDNA in 5´-3´ direction binds to its complementary ssDNA only if this complementary sequence is in 3-5 direction. But, as shown in Figure 6.8, this parallel binding also produces very intense florescence signals. It becomes very difficult to distinguish these false signals through existing statistical analysis and bioinformatics tools.

We use modified southern blotting for hybridization. Here ssDNA which is extracted from our samples and represents relevant data is immobilized on a membrane/surface. Then complementary ssDNA sequences labelled with Dig-label are hybridized with the immobilized targets. The blotting experiments, where our S.I.S. are positively compared with same S.I.S ordered from Metabion is presented in Section 6.7. The data produced from our techniques highly corresponds to desired results. Due, to constraints on existing commercial sequencing platforms, it is difficult to directly sequence ssDNA of 40 nt length. So this ssDNA is again inserted into a dsDNA sequence of length 135 bp. This is accomplished through molecular cloning. The molecular cloning procedure ans subsequent Sanger sequencing are shown in Section 6.8. The same procedure used to build the Toy system, is used to clone the S.I.S. into PCR1.2 vector from Fermentas GmbH. This vector is transformed into E.Coli cells and the resulting colonies are shown in Figure 6.10. The 135 bp sequence is extracted from the colonies through colony PCR. This is shown in Figure 6.11. Then these 135 bp sequences are sent for Sanger sequencing to Gatc

Biotech AG.

Sanger sequencing is performed on results presented in Section 6.8, show that the data from our ssDNA sequences produced from our MEA technique is the specific. This means, the information from the sequencing data reveals exactly the 40 nt ssDNA that was embedded in the complex DNA mixture in section.

7. Discussion

Data is dirty

Information is gold.

Previous research has produced evolved experimental procedures which produce increased amounts of raw data from various enzyme based fragmentations (Ranade et al., 2009), (Klur et al., 2004), PCR and variations of PCR based techniques (Zhang et al., 1998). However, raw data is not actual information. Data is a collection of unprocessed measurements and information is processed data. All raw data is not useful.

The goal of this dissertation is to show the importance of information, which even if generated in reduced amounts, is much more viable and accurate than generation of information from large amounts of data with various statistical models and various commercial platforms (Prodromou et al., 2007).

In the following sections, identification, retrieval and validation of specific information are discussed with help of the experimental results from various molecular biological methods. In the end of the section, a conclusion is presented to explain the importance of our Toy system approach, use of new methods for accurate information retrieval and purification.

7.1. Holistic approach

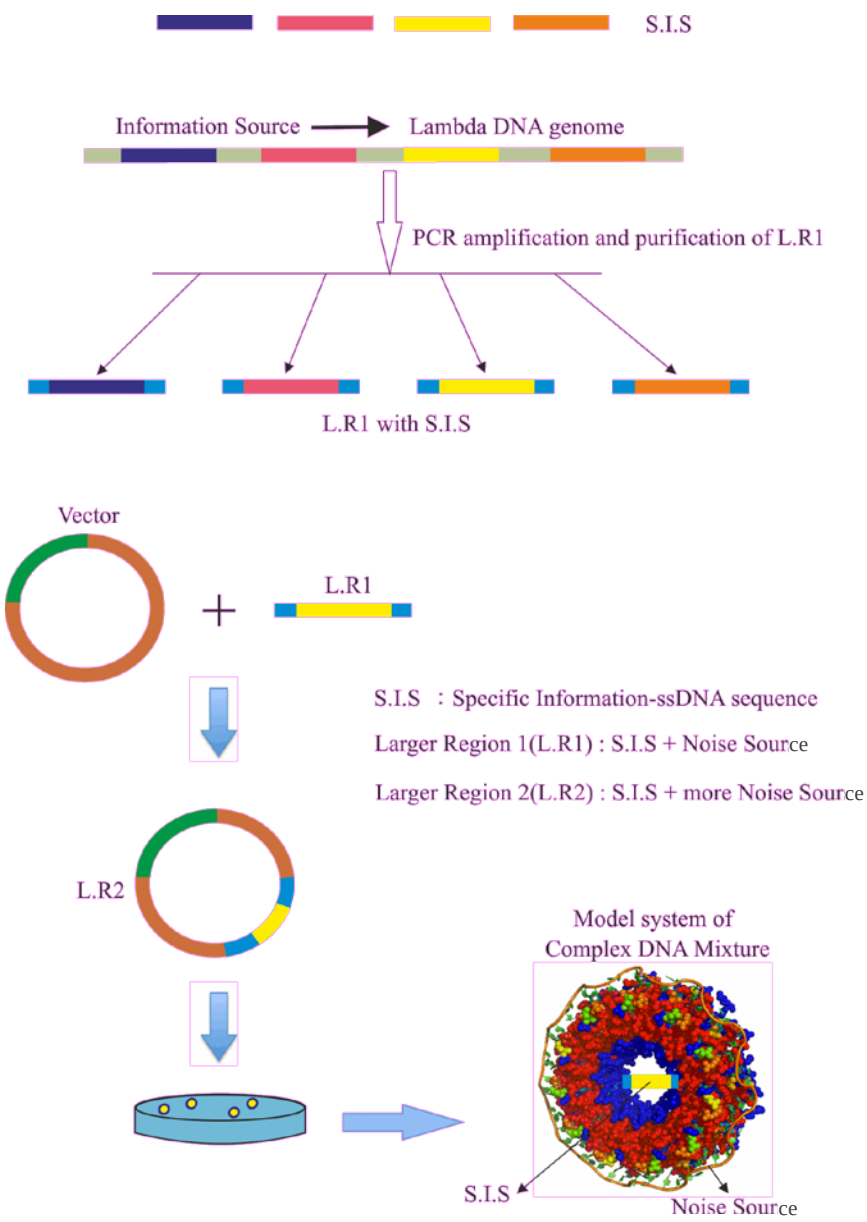


Figure 7.1: Holistic approach for complex DNA mixture-Stepwise insertion of specific information into a noise source

In the Figure 7.1, Specific Information sequences (S.I.S.) are isolated from an information source (Lambda DNA genome). This is accomplished by PCR based amplification of a larger stretch/region of DNA from the lambda DNA genome. This larger region (L.R.1) contains the S.I.S. This step results in addition of unspecific hybridization sources fthat may

result in experimental noise. Then, larger region (L.R.1) is ligated to a linearized vector through molecular cloning to create a circular vector (L.R.2) that can be transformed into bacterial cells. After successful transformation and a complex DNA mixture is obtained. It can be clearly understood that specific information S.I.S is gradually clouded with unspecific information or noise sources. Through this holistic approach, the complexity of the model system can be controlled. This model results in understanding the constraints on accurate information transfer.

7.2. Reductionist approach

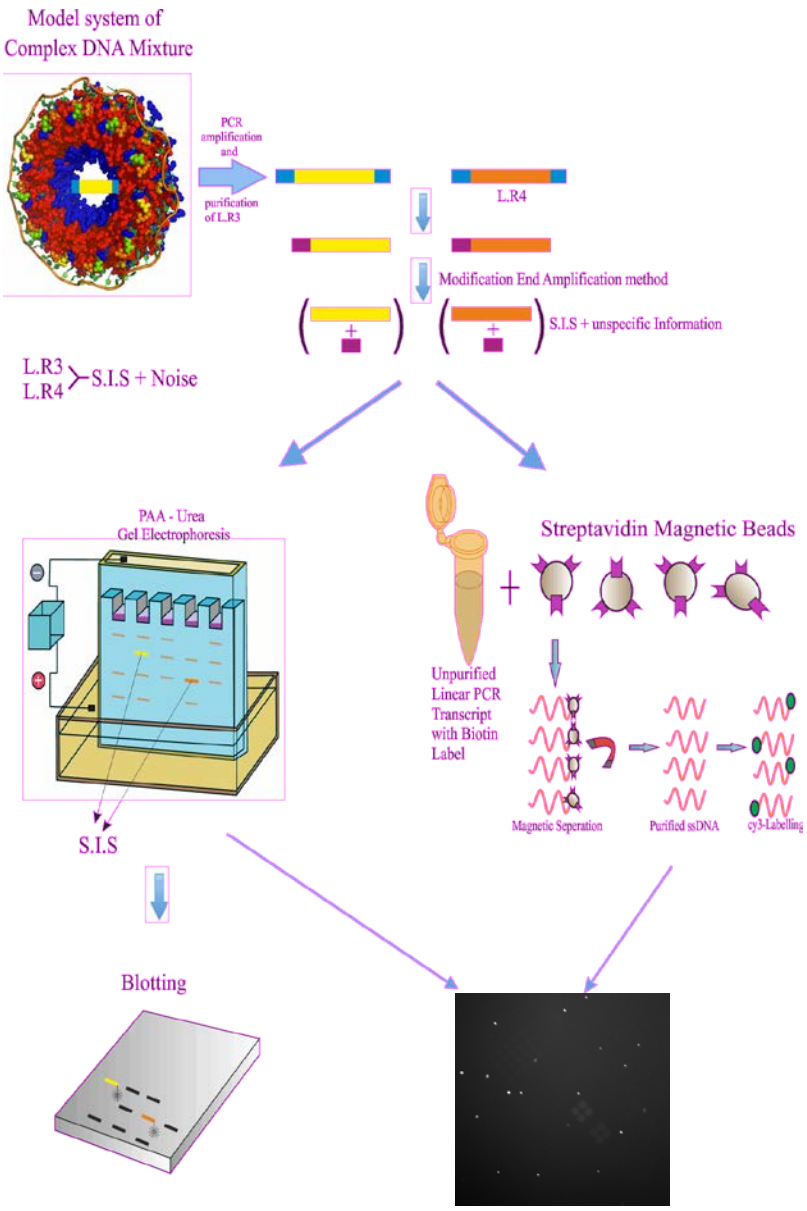


Figure 7.2: Reductionist approach for information retrieval-Stepwise retrieval and verification of specific information

In Figure 7.2, the reductionist model is presented. From the complex DNA mixture, again larger stretches of DNA (L.R.3 and L.R.4) containing S.I.S. and unspecific binding regions are isolated through PCR based amplification. Then, our novel M.E.A. technique is applied to linearly amplify S.I.S. from L.R.3 and L.R.4. The product of M.E.A. technique is verified

for the presence of S.I.S. through PAA-Urea gel electrophoresis. Then, S.I.S. is extracted/purified through nucleic acid isolation and purification techniques like Biotin-Streptavidin magnetic beads procedure. This purified sequence can be labelled with any type of fluorescent dye for further analysis. The fidelity of the S.I.S. is further investigated with membrane/surface based hybridization techniques like DNA microarrays and Southern blotting. It can be clearly understood that, the retrieval of specific information from complex DNA mixture is stepwise. In every step, the extracted data is validated to check if it relates the intended specific information sequences. This results in an approach where the analysis through various hybridization techniques becomes more reliable. This enables this experimental model based approach to be platform/software independent and more reproducible.

7.3. Information overflow

In this dissertation, short stretches of ssDNA which have a length between 20-100 nt, are identified first and then they are inserted into a Toy system. These are referred to as specific information sequences S.I.S. Incorporation of this S.I.S. into a noise environment and subsequent retrieval represents the basis of this dissertation. This is different from existing and conventional fragmentation procedures in the following ways.

1. **A predefined hypothesis results in emphasis on generating and validating a limited amount of information**

   Emphasis on particular set of experiments which generate only specific data is important to validate a hypothesis (Gibson, 2003). Subsequently, there is no comparison between two samples or classes to prove an assumption that information flow is accurate or inaccurate. There is always a possibility that the experimental condition determines or

adversely influences the outcome of the experiments. Sometimes, these conditions result in wrong class identification (Tarca et al., 2006).

2. High dependence on data processing algorithms

Previous research has shown that genome based high throughput analysis depends heavily on pre-processing of microarray data, background corrections for unspecific hybridization and cross hybridization (Sifakis et al., 2012; Tarca et al., 2006). Here the data is obtained from various fragmentation techniques. So the unspecific hybridization mostly arises from sample preparation as large amounts of fragments result in unspecific hybridization. It becomes difficult to separate specific information from raw data.

3. Linear PCR and LATE PCR for ssDNA generation

Most of the existing methods provide reliable results, like the restriction enzymes and endonucleases used for fragment library generation and cleaving of DNA sequences at particular regions. Then large amounts of fragments which are generated through these diverse techniques are analysed with high throughput analysis (Ranade et al., 2009; Saupe et al., 1998). Subsequently, increased number of fragments provides large amounts of data which contains unspecific lost information.

Most of these techniques produce fragments of DNA/RNA which has no relevance for particular application. Also, they produce ssDNA sequences which have lengths more than 200 nt.

4. **Profound influence of unspecific binding on noise generation**

Many gene expression profiling techniques depend entirely on preferential binding for investigation of active genes and specific regions of genome. It is very important to reduce the influence of unspecific binding as this leads to wrong analysis or partial analysis.

5. **Problems with longer ssDNA sequences for highthoroughput analysis**

The inherent properties of ssDNA compel the DNA strands to form loops. This results in more thermodynamic stability. However, this results in unstable and nonspecific binding. Consequently, there is an irreversible loss of valuable information due to the nucleotides involved in loop formation. Moreover, this loop and hairpin forming tendency increases with the length of ssDNA sequences (Trapp et al., 2011).

7.4. Importance of novel experimental model based analysis

The novel MEA method used to identify and amplify specific ssDNA sequences that represent specific information provides a solution to separate noise sources during the initial step of information retrieval. This method functions through the control of fundamental biomolecular interactions like DNA-DNA hybridization for information transfer instead of random fragmentation. The retrieved data is screened through various molecular biology and biochemistry based techniques like gel electrophoresis and HPLC. Then the information is purified for removal of possible noise sources and subsequently verified through blotting for successful information transfer and further through Sanger

sequencing for verifying the nucleotide composition. This can significantly reduce the problems with conventional techniques by generating reduced amounts of short specific information sequences.

SAGE (Serial analysis of gene expression) is a similar model for isolation of short transcripts from a cDNA sequence (Matsumura et al., 2005). These short transcripts are ligated together and cloned into a vector for sequencing. SAGE is a much more accurate and quantitative analysis tool in comparison with microarrays. But, due to use of fragmentation methods, there is a redundancy in the short transcripts that are generated. So, the information can occur more than once in the final sequence analysis. This problem is eliminated in our novel MEA technique due to the use of enzymes for linear amplification. Here the presence of enzyme is the limiting factor. The enzymes used in the MEA technique need a primer binding site in the complex DNA mixture to generate linearly amplified ssDNA sequence/specific information. This means that information can be generated only when a primer binds to its specific binding site and only then can the enzyme create an amplified sequence through addition on nucleotides. This results in reduction of unspecific binding and generation of unspecific information.

When multiple primers are used with MEA technique, multiple information sequences are generated simultaneously. Due to presence of multiple primers, there is a possibility of cross binding between the primers. Also, some of the primers may have a degree of similarity in the nucleotide composition. This can result in unspecific binding of the primers to regions in the information source/complex DNA mixture. This results in generation of unspecific information and information overflow.

This MEA technique can be generalized by using a single primer sequence (universal

primer) that can binding to many regions in a complex DNA mixture. This reduces the information overflow generated by the presence of multiple primers. This universal primer should be specific for the complex DNA mixture.

7.5. Relevance to biological information systems

The adverse effects of irreversible loss information during signal transmission and retrieval has been a major problem in data communication systems. The ability of the transmitting system to reliably transmit data is limited by channel capacity of the system. The importance of channel capacity is defined in the information theory developed by Shannon (Shannon, 1948). There have been many advances in increasing the accuracy of communication systems through novel electronic circuitry. But, channel capacity still defines the uppermost limit for accurate information transmission.

Channel capacity specifies the maximum amount of reliable information that can be accurately transmitted. This theory can also be applied to biological information analysis. Information theory developed by Shannon has been applied and proved to be valid for information transfer in biological systems. This is referred to as molecular information theory and was proposed among others, by Tom Schneider (Schneider, 2000). This theory is a hypothesis for relation between information theory and binding processes in genetic systems. The evolution of biological systems is highly dependent on the information gained by genetic systems. High and accurate information gain results in a highly evolved biological system. The uncertainty in information retrieved from biological systems can be attributed to changes in genetic framework of the system. This uncertainty can result from changing population dynamics of biological systems (Rivoire and Leibler, 2010).

The amount of specific information that can be accurately validated by hybridization

based sequence analysis methods depends on accurate information retrieval. Our experimental approach can be very important in creating an upper limit for channel capacity in information theory (Shannon, 1948). Through of MEA technique, it is possible to reduce the irreversible loss of information and reduce the addition of noise. This has a direct effect on avoiding information overflow.

8. Conclusions

Most of the techniques used for the information transfer analysis tend to prefer ***brute-force*** method. They produce randomly large amounts of fragments from any given information source of interest.

This has the following problems:

- Large amounts of unspecific data ***cloud*** specific data and specific information is irreversibly lost.
- A fragmented mixture is generated which has the following components

Fragmented mixture=Specific information+Unspecific information+Mixed information

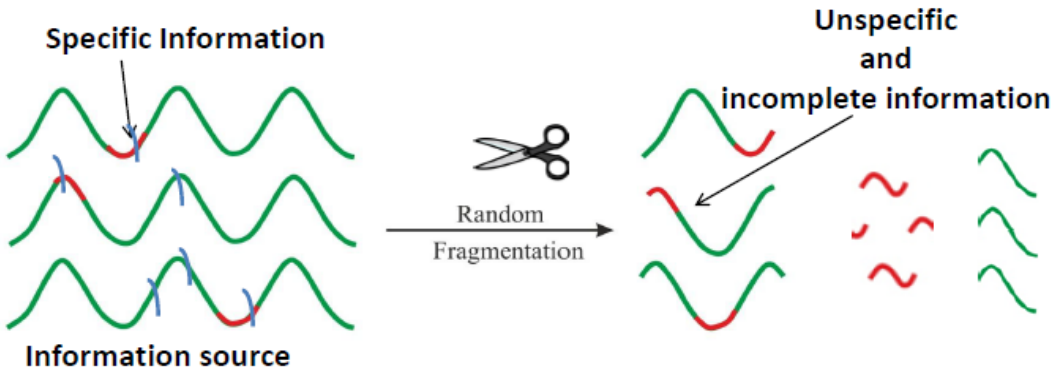


Figure 8.1: Gene expression profiling

To remedy the above problems through our work, initial direct approaches to retrieve specific information from a complex DNA mixture were tried. These approaches did not yield conclusive and repeatable results. Then a stepwise methodology was developed and applied to insert specific information into a noise source. Then this information is retrieved

using our novel MEA technique and verified with hybridization experiments.

In this dissertation, we propose and validate a model to solve the fundamental problems that hamper the enormous potential of gene expression profiling.

When analysing information flow from DNA to protein, generating specific information from particular stretches of DNA makes the experimental data very accurate. This can be either a gene or a cluster of genes. By concentrating on specific regions and generating multiple shorter ssDNA sequences, which have a predefined length and nucleotide composition, the following discrepancies are avoided.

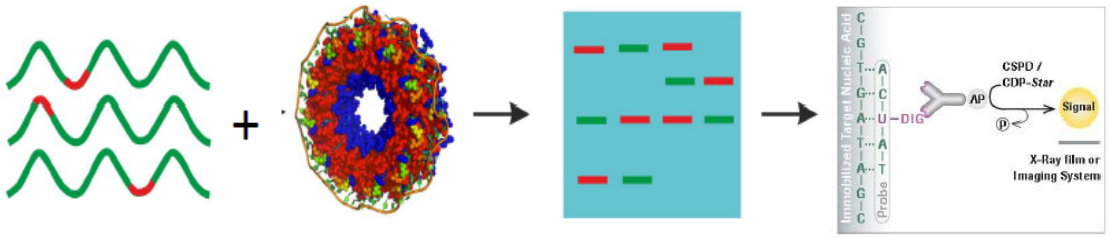
- Longer fragments form loops and hairpin structures. So part of the sequences do not hybridize with their complementary sequences and result in information loss.
- Generating reduced amounts of raw data for analysis eases the significant strain on the capabilities of microarray analysis and reduces platform dependency.
- By generating an information mixture, dead-data, which is data that has no use for the particular experiment or hypothesis, is produced thereby giving unspecific signals and information which clouds the specific information.
- Dynamic equilibrium is much more difficult to reach when more sequences compete to hybridize with the sequences on microarrays.
- When a portion of dead data has some stretch of ssDNA that is similar to the immobilized sequences, the possibility of unspecific or partial hybridization is higher.
- The major constraint of high throughput applications is the cost aspect. This can be reduced by fabricating microarrays with specific probes instead of large amounts of sequences, which may or may not be important for the particular experiment/hypothesis.

Need for an experimental system based approach

Through holistic approach, specific information is systematically embedded into a noise environment that provides many possibilities for incorrect information transfer.

- Unspecific hybridization
- Production of large amounts of dead data
- Synthesis of longer sequences which form loops
- Synthesis of a fragmented mixture that has dead data, incomplete information produced from existing fragmentation techniques
- Strain on high throughput applications through large data processing requirements.
- Possible introduction of intrinsic and extrinsic noise sources and random mutations through introduction of bacterial culture as the final step to express our specific information inside a cell.

By using a reductionist approach for retrieval of specific information, the possibility of incomplete and defective information retrieval is introduced. All the noise sources that were introduced through holistic approach can influence the identification, isolation and retrieval of accurate information from the Toy system. Thus, the shortcomings of existing information retrieval and analysis techniques are eliminated.



Specific Information (Red) Noise Retrieval Verification

Figure: 8.2 Reductionist Approach for retrieval and verification of specific information

Our Toy system is a complex DNA mixture which has noise and specific information that is realized through various molecular biological and biochemical techniques. Using this Toy system as basis, our MEA methods retrieve reduced amounts of information which is highly specific in nature as opposed to existing techniques is established. Figure 8.2 shows the reductionist approach for step-wise information retrieval and verification from complex DNA mixture. This model is dependent on the biomolecular interactions that govern the information transfer in various processes like molecular binding; DNA replication and hybridization are understood.

The intention behind choosing a systematic model based approach is to replicate the possibilities of noise sources and loss of specific information. Instead of comparing one set of data to another at amount of specific information. Successful retrieval of this specific information depends on the ability to remove the noise sources at fundamental level.

9. Summary and outlook

A stepwise methodology was applied to improve the fidelity of information retrieval and verification. Initially, specific information is inserted into a noise source. Then this specific information is retrieved and verified with Southern blotting-Hybridization and Sanger sequencing. This dissertation proposes a model biomolecular system where specific information was clouded by noise sources. This is a complex DNA mixture that represents interactions between biomolecules at a fundamental level. Then the specific information was retrieved from this model biomolecular system. We show that instead of extracting higher number random sequences from a template, it is possible to identify and retrieve specific sequences of predefined length and nucleotide composition. It is also possible to apply our model to systems of higher complexities. Our experiments also show higher degree of reproducible results and mostly specific data for analysis. This adds more specificity to our model and shows that specific ssDNA sequences are extracted from a system that has noise sources of various complexities.

The findings in this dissertation provide an excellent starting point for many important applications. The novel sample preparation methods proposed in this dissertation can be extended to conventional oligonucleotide microarray analysis. It will enable the microarray analysis to be more independent of commercial platforms and possibly reduce the discrepancies in gene expression studies. Also, it is much viable to formulate experimental model based strategies to investigate and identify problems in genetic information transfer like mutations and single nucleotide polymorphism. Thus, accurate identification of particular regions of interest is made possible. This eliminates more dependency on statistical algorithms. The sample preparation techniques can be applied to commercial lab-on-chip applications like chip based sequencing and chip based disease identification. These

techniques will remove probable noise sources present in the biological sample and thus make hand-held sequencing platforms more accessible and reliable. The removal of noise sources is very important in any form of gene expression or high throughput studies.

It is easier to understand constraints with information transfer with analysis based on experimental procedures. This experimental approach differs from many existing theoretical models in the sense that the specific information is extracted based on basic biomolecular interactions like DNA-DNA binding and dependence on temperature for hybridization. Our model proposes a work flow to understand information overflow and the experimental results like gel electrophoresis and HPLC provide adequate proof as opposed to theoretical and mathematical models. Problems resulting from incorrect synthesis of microarrays show that experimental noise is highly dependent on correct synthesis and subsequent error-free analysis. By applying a reductionist approach for stepwise retrieval and validation of information, many factors that result in generation of unspecific information and subsequent error strewn analysis can be reduced greatly.

The proposed MEA technique addresses information overflow and irreversible loss of information through a model/toy system. The connection between Shannon information theory and MEA technique should be further investigated by using a real biological system as model for genetic information transfer and theory.

List of Figures

List of Tables

10. References

[1]. **Crick, F.** (1970). Central Dogma of Molecular Biology. *Nature* **227**.

[2]. **Draminski, M., Rada-Iglesias, A., Enroth, S., Wadelius, C., Koronacki, J. and Komorowski, J.** (2008). Monte Carlo feature selection for supervised classification. *Bioinformatics (Oxford, England)* **24**, 110–7.

[3]. **Fukano, H. and Suzuki, Y.** (2009). Enzymatic conversion of long DNA to small DNA fragments for the construction of short hairpin RNA expression libraries. *Analytical biochemistry* **385**, 80–4.

[4]. **Gibson, G.** (2003). Microarray analysis: genome-scale hypothesis scanning. *PLoS biology* **1**, E15.

[5]. **Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E. M., Sipos, B. and Birney, E.** (2013). Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **494**, 77–80.

[6]. **Graham, D. L., Ferreira, H. a, Freitas, P. P. and Cabral, J. M. S.** (2003). High sensitivity detection of molecular recognition using magnetically labelled biomolecules and magnetoresistive sensors. *Biosensors & bioelectronics* **18**, 483–8.

[7]. **Greiner, O. and Day, P. J. .** (2004). Avoidance of nonspecific hybridization by employing oligonucleotide micro-arrays generated from hydrolysis polymerase chain reaction probe sequences. *Analytical Biochemistry* **324**, 197–203.

[8]. **Hacker,D.L., Petty,I.T.D., Wei,N., Morris, T.J.** (1992). Turnip crinkle virus genes required for RNA replication and virus movement, Virology **186**, 1-8.

[9]. **Klur, S., Toy, K., Williams, M. P. and Certa, U.** (2004). Evaluation of procedures for amplification of small-size samples for hybridization on microarrays. *Genomics* **83**, 508–17.

[10]. **Kobayashi, T., Mikami, S., Yokoyama, S. and Imataka, H.** (2007). An improved cell-free system for picornavirus synthesis. *Journal of virological methods* **142**, 182–8.

[11]. **Kuznetsov, S. V., Ren, C.-C., Woodson, S. a and Ansari, A.** (2008). Loop dependence of the stability and dynamics of nucleic acid hairpins. *Nucleic acids research* **36**, 1098–112.

[12]. **Lindow, M., Vornlocher, H.-P., Riley, D., Kornbrust, D. J., Burchard, J., Whiteley, L. O., Kamens, J., Thompson, J. D., Nochur, S., Younis, H.** (2012). Assessing unintended hybridization-induced biological effects of oligonucleotides. *Nature biotechnology* **30**, 920–3.

[13]. **Mardis, E. R.** (2008). The impact of next-generation sequencing technology on genetics. *Trends in genetics : TIG* **24**, 133–41.

[14]. **Missiuro, P. V., Liu, K., Zou, L., Ross, B. C., Zhao, G., Liu, J. S. and Ge, H.** (2009). Information flow analysis of interactome networks. *PLoS computational biology* **5**, e1000350.

[15]. **Matsumura, H., Ito, A., Saitoh, H., Winter, P., Kahl, G., Reuter, M., Krüger, D. H. and Terauchi, R.** (2005). SuperSAGE. *Cellular microbiology* **7**, 11–8.

[16]. **Nordstro, T. and Alderborn, A.** (2002). Method for one-step preparation of double-stranded DNA template applicable for use with Pyrosequencing technology. Journal of Biochemical and Biophysical Methods**, 52 (2),** 71-82.

[17]. **Orgel, L. E.** (2004). Prebiotic chemistry and the origin of the RNA world. *Critical reviews in biochemistry and molecular biology* **39**, 99–123.

[18]. **Pavlov, A. R., Pavlova, N. V, Kozyavkin, S. a and Slesarev, A. I.** (2004). Recent developments in the optimization of thermostable DNA polymerases for efficient applications. *Trends in biotechnology* **22**, 253–60.

[19]. **Petersen, K., Oyan, A. M., Rostad, K., Olsen, S., Bø, T. H., Salvesen, H. B., Gjertsen, B. T., Bruserud, O., Halvorsen, O. J., Akslen, L. A., et al.** (2007). Comparison of nucleic acid targets prepared from total RNA or poly(A) RNA for DNA oligonucleotide microarray hybridization. *Analytical biochemistry* **366**, 46–58.

[20]. **Pozhitkov, A. E., Nies, G., Kleinhenz, B., Tautz, D. and Noble, P. a** (2008). Simultaneous quantification of multiple nucleic acid targets in complex rRNA mixtures using high density microarrays and nonspecific hybridization as a source of information. *Journal of microbiological methods* **75**, 92–102.

[21]. **Prodromou, C., Savva, R. and Driscoll, P. C.** (2007). DNA fragmentation-based combinatorial approaches to soluble protein expression Part I. Generating DNA fragment libraries. *Drug discovery today* **12**, 931–8.

[22]. **Ranade, S. S., Chung, C. B., Zon, G. and Boyd, V. L.** (2009). Preparation of genome-wide DNA fragment libraries using bisulfite in polyacrylamide gel electrophoresis slices with formamide denaturation and quality control for massively parallel sequencing by oligonucleotide ligation and detection. *Analytical biochemistry* **390**, 126–35.

[23]. **Riva, A., Carpentier, A.-S., Torrésani, B. and Hénaut, A.** (2005). Comments on selected fundamental aspects of microarray analysis. *Computational biology and chemistry* **29**, 319–36.

[24]. **Rivoire, O. and Leibler, S.** (2010). The Value of Information for Populations in Varying Environments. *Journal of Statistical Physics,* **142**, Issue 6, pp.1124-1166.

[25]. **Sanchez, J. A., Pierce, K. E., Rice, J. E. and Wangh, L. J.** (2004). Linear-after-the-exponential (LATE)-PCR: an advanced method of asymmetric PCR and its uses in quantitative real-time analysis. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 1933–8.

[26]. **Sanger, F. and Nicklen, S.** (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 5463–5467.

[27]. **Saupe, S., Bernard, P., Laurent-Brun, E., Derancourt, J. and Roizès, G.** (1998). Construction of a human BcgI DNA fragment library. *Gene* **213**, 17–22.

[28]. **Schena, M., Shalon, D., Davis, R. W. and Brown, P. O.** (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)* **270**, 467–70.

[29]. **Schneider, T. D.** (2000). Evolution of biological information. *Nucleic acids research* **28**, 2794–9.

[30]. **Seliger, H., Hinz, M. and Happ, E.** (2003). Arrays of immobilized oligonucleotides--contributions to nucleic acids technology. *Current pharmaceutical biotechnology* **4**, 379–95.

[31]. **Shannon, C. E.** (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal* **27**, 379–423, 623–656.

[32]. **Sifakis, E. G., Prentza, a, Koutsouris, D. and Chatziioannou, a a** (2012). Evaluating the effect of various background correction methods regarding noise reduction, in two-channel microarray data. *Computers in biology and medicine* **42**, 19–29.

[33]. **Southern, E. M.** (1992). Detection of specific sequences among DNA fragments separated by gel electrophoresis. 1975. *Biotechnology (Reading, Mass.)* **24**, 122–39.

[34]. **Stenlund, A., Perricaudet, M., Tiollais, P. and Pettersson, U.** (1980). of restriction enzyme fragment. Gene, **10 (1)**, 47-52.

[35]. **Syvänen, C.** (1999). From gels to chips: “minisequencing” primer extension for analysis of point mutations and single nucleotide polymorphisms. *Human mutation* **13**, 1–10.

[36]. **Tan, P. K., Downey, T. J., Jr, E. L. S., Xu, P., Fu, D., Dimitrov, D. S., Lempicki, R. A., Raaka, B. M. and Cam, M. C.** (2003). Evaluation of gene expression measurements from commercial microarray platforms. **31**, 5676–5684.

[37]. **Tarca, A. L., Romero, R. and Draghici, S.** (2006). Analysis of microarray experiments of gene expression profiling. *American journal of obstetrics and gynecology* **195**, 373–88.

[38]. **Vincent, M., Xu, Y. and Kong, H.** (2004). Helicase-dependent isothermal DNA amplification. *EMBO reports* **5**, 795–800.

[39]. **Wang, J., Cai, X., Rivas, G., Shiraishi, H. and Dontha, N.** (1997). recognition and detection at chronopotentiometric DNA chips *. **12**, 587–599.

[40]. **Weissman, S. M.** (1979). Current approaches to analysis of the nucleotide sequence of DNA. *Analytical biochemistry* **98**, 243–53.

[41]. **Wen, D. and Zhang, C.** (2012). Universal Multiplex PCR: a novel method of simultaneous amplification of multiple DNA fragments. *Plant methods* **8**, 32.

[42]. **Zhang, D. Y., Brandwein, M., Hsuih, T. C. and Li, H.** (1998). Amplification of target-specific, ligation-dependent circular probe. *Gene* **211**, 277–85.

[43]. **Zhang, D. Y., Zhang, W., Li, X. and Konomi, Y.** (2001). Detection of rare DNA targets by isothermal ramification amplification. *Gene* **274**, 209–16.