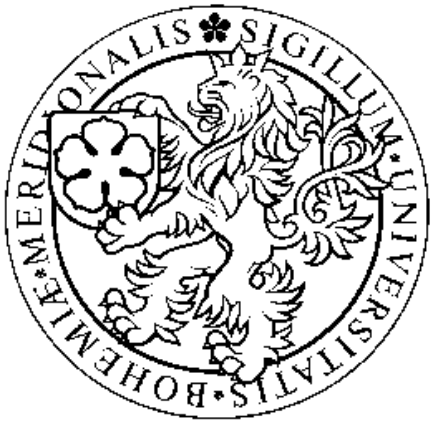
University of South Bohemia

Faculty of Science



DNA Separation by Capillary Electrophoresis: A Literature Overview

Author: Daniel Gstöttenmayr

Supervisor: Alexander W. Bruce *Ph.D.*

December, 2010

Annotation

This thesis is a literature overview of the application of Capillary Electrophoresis (CE) in the analysis of genes and genomes. In the Introduction the question of why gene analysis is important and the advantages of the application of CE in this field are addressed. Also a brief overview of the separation of deoxyribonucleic acid (DNA) by techniques other than CE is given. Further a summary about the currently available reviews dealing with the applications of CE in DNA diagnostics is included. In the theory chapters the fundamentals of DNA are covered. Further a brief outline of the principles of CE essential for applications in DNA analysis is included. In the second half a survey of the CE methodologies for genome sequencing is presented.

Im Zuge dieser Bachelorarbeit wurde eine Literaturübersicht über die Analyse von Desoxyribonukleinsäure (DNA) mittels Kapillarelektrophorese (CE) erarbeitet. In der Einleitung werden die Bedeutung von Gen Analysen und die Vorteile von CE in diesem Bereich näher erläutert. Auch wird ein kurzer Überblick über andere Techniken, welche sich ebenfalls zur DNA Auftrennung eignen, gegeben. In einer kurzen Zusammenfassung werden bereits bestehende Arbeiten, welche sich mit DNA Analysen mittels CE beschäftigten, behandelt. Die Theorie Kapiteln setzen sich mit den Grundlagen der CE und den Grundlagen zum Thema DNA auseinander. Weiters werden Prinzipien der CE, welche essentiell für die Anwendung in der DNA Analyse sind, behandelt. In der zweiten Hälfte dieser Arbeit werden Verfahren zur DNA Analyse mittels CE näher erörtert.

Affirmation

I hereby declare under oath that the submitted bachelor thesis has been written solely by me without any third-party assistance. Additional sources or aids are fully documented in this paper, and sources for literal or paraphrased quotes are accurately credited.

I hereby declare that, in accordance with Article 47b of Act No. 111/1998 in the valid wording, I agree with the publication of my bachelor thesis, in full to be kept in the Faculty of Science archive, in electronic form in publicly accessible part of the STAG database operated by the University of South Bohemia in České Budějovice accessible through its web pages. Further, I agree to the electronic publication of the comments of my supervisor and thesis opponents and the record of the proceedings and results of the thesis defence in accordance with aforementioned Act No. 111/1998. I also agree to the comparison of the text of my thesis with the Teses.cz thesis database operated by the National Registry of University Theses and a plagerism detection system.

Date ……

Signature …….

Acknowledgement

I am grateful to Alexander W. Bruce *Ph.D.* for supervising my Bachelor thesis. Further I want to thank my family, friends and colleagues for supporting me.

List of Abbreviations

A	adenine
AFLP	amplified fragment length polymorphism
BGE	background electrolyte
bp	base pair
C	cytosine
CAE	capillary array electrophoresis
CE	capillary electrophoresis
DNA	deoxyribonucleic acid
dsDNA	double stranded deoxyribonucleic acid
EOF	electroosmotic flow
G	guanine
HGP	human genome project
LIF	laser induced fluorescence
miRNA	micro ribonucleic acid
MS	mass spectrometry
PCR	polymerase chain reaction
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
SDS	sodium-dodecyl sulfate
SGE	slab gel electrophoresis
siRNA	small interfering ribonucleic acid
SNP	single nucleotide polymorphism
SSCP	single-strand conformation polymorphism
ssDNA	single stranded deoxyribonucleic acid
T	thymine
U	uracil
UV	ultra violet

Table of Contents

1 Introduction

The tools and techniques of molecular biology are essential components to understanding how organisms work. For example, a successful genetic analysis provides information that can be used in clinical genetic diagnostics. Even though most diseases must be viewed in the larger context of a cell or an organism it is unquestionable that the development and investigation of reliable cost effective high-throughput analytical technologies are of major importance.

During the life of a living organism its DNA undergoes changes which can be divided in two groups. There are DNA mutations and polymorphism, each of which are stable changes due to damage or not corrected alterations. Polymorphisms occur in at least one percent of the population and are not harmful, whereas mutations are present in less than one percent and can result in disease or increased risk of developing a disease. DNA can undergo several types of changes: point mutations, which are substitutions of a single nucleotide; deletions or insertion of a single nucleotide; deletion or insertion of a group of nucleotides. When we speak about alterations in the coding regions, the deletion or insertion of a single nucleotide can result in a reading frame shift of the codon sequence. Since every amino acid is coded by three nucleotides, the reading frame shift can lead to the translation of no protein at all or the formation of an altered protein. If the alterations occur in a non coding region the protein sequence is not directly affected but the protein synthesis is likely to be affected resulting in defects related to quantity or spatio-temporal expression. Among the group of polymorphisms there is a type that is called tandem repeat polymorphism. This type can serve as an individual's barcode for medical and forensic purposes. Another type is the single nucleotide polymorphism (SNP). This polymorphism provides information that can be used as a genetic marker, too. Using SNPs it may be possible to predict the genetic risk of developing a certain disease, to be able to diagnose a disease more accurately or even to predict the therapeutic response to a drug. Utimately the goal is to develop therapies that consider the individual genetic profile of individual patients, rather than populations of various SNPs. SNPs may possibly provide closer insights into the biological evolution of related organisms. The investigation of all these complex genetic problems requires the analysis of SNPs on a large scale. In order to be able to perform these measurements high throughput analytical methods, which can be automated are necessary, are required.

One method to determine polymorphisms, mutations, or the DNA sequence is capillary electrophoresis (CE). Within this method DNA fragments are separated in an electrical field according their differing mobilites. The capillary, where the separation is carried out, contains a sieving matrix, thus the fragments can be separated according their differing lengths. The

differences in fragment lengths can be used either, to detect polymorphisms and mutations, or to determine the sequence. In order to do genetic analysis on a large scale high throughput analytical methods are required. CE fulfils the requirements, since it is suitable for high throughput analysis and it addresses the problem of amenability to automation. Besides this, it is a method with low operation costs, which possesses high speed separation capabilities. A further reason for the employment of CE in the field of genetic analysis is the ease of use and the fact that it offers a high overall separation performance.

There are other techniques that can serve as alternatives to CE. These include, hybridization on DNA microarrays [1], *in situ* hybridization [2], denaturing liquid chromatography [3], mass spectrometry [4], flow cytometry [5], quantitative or real time polymerase chain reaction (PCR) [6], pyrosequencing [7] and single-molecule sequencing [8].

The applications of CE in genetic analysis have been reviewed in several articles. Klepárník and Boček published a comprehensive overview and literature survey on the DNA diagnostics by CE [9]. In a review written by Keith R. Mitchelson the use of CE for DNA polymorphism analysis was described [10]. Slater *et al.*, published an article on the theory of DNA separation by capillary electrophoresis [11].

This thesis is aimed to provide an insight into the various principles employed in genetic diagnostics and it should further contribute to a deeper understanding of the possibilities that CE offers in the area of gene analysis.

2 Deoxyribonucleic Acid (DNA)

2.1 Principles

The key concept of DNA is that the genetic information is “stored” within the sequence of the nucleic acid. The DNA provides the genetic information, which is required to construct and maintain a copy of a living organism. Therefore it is the material of heredity in most organisms. The central dogma in molecular biology is: DNA makes RNA makes proteins. First DNA is transcribed in order to get a RNA copy. Then, by the process of translation, the RNA is decoded to produce proteins. Although it is now becoming clearer that RNAs themselves, in addition to proteins, play many fundamental roles during gene expression and readout of the DNA genome *e.g.* microRNAs (miRNAs) and small interfering RNAs (siRNAs).

2.2 Structure of Nucleic Acids

DNA is a double-stranded antiparallel helix which is held together by hydrogen bonding between chemical moieties, commonly referred to as base pairing. The DNA is a polymeric molecule made up of linear chains. The structure of DNA and RNA is illustrated in Figure 2.1. Each chain is made up of subunits, which are called nucleotides. A nucleotide can be divided into three parts: a nitrogenous base, a five-carbon-atom, pentose, sugar and a phosphate group. Together the base and the pentose sugar form a so called nucleoside. The whole subunit, consisting of base, sugar and phosphate, is called the nucleotide monophosphate or simply a nucleotide. The base can be a purine or pyrimidine base. DNA and RNA contain the same purines adenine (A) and guanine (G) and the same pyrimidine cytosine (C). The pyrimidine thymine (T) is only found in DNA, while RNA contains the pyrimidine uracil (U). RNA contains the sugar ribose, thus its nucleotides are termed ribonucleotides. DNA contains a ribose which lacks a hydroxyl group at the 2' carbon. This sugar is known as 2'-deoxyribose, so the DNA nucleotides are termed deoxyribonucleotides. The pentose sugar is connected via its 1' carbon to the nitrogen of the base. Individual nucleotides of RNA and DNA are connected via sugar phosphate, or phospho-diester, bonds. This is a linkage between the 3'carbon of one nucleotide and the phosphate group of the 5'carbon of the other nucleotide (see Figure 2.2). Two nucleotides connected together are called a dinucleotide, three nucleotides are called trinucleotides and so forth. Multiple nucleotides are termed oligonucleotides or polynucleotides, depending upon their respective lengths. As already mentioned above, the DNA is a double stranded anti-parallel helix made up of the two polynucleotide chains. Each strand has a 5' phosphate end and a 3' hydroxyl end. One helix chain runs in 3' to 5' orientation, the other one in 5' to 3' orientation relative to the first strand. Thus, the pairing of a purine (A or G) with a pyrimidine (T or C) via hydrogen bonds leads to the formation of the anti-parallel double stranded DNA. Furthermore, an A base on one strand always base pairs with a T on the opposite strand and *vice-versa*. The same relationship is true between G and C base pairs, hence any double-stranded DNA molecule always contains a 1:1 ratio of A & T and G & C nucleotides. This relationship is known as Chargaffs rule after it's discoverer. The structure of the nucleotide chains and the base pairing is illustrated in Figure 2.1.

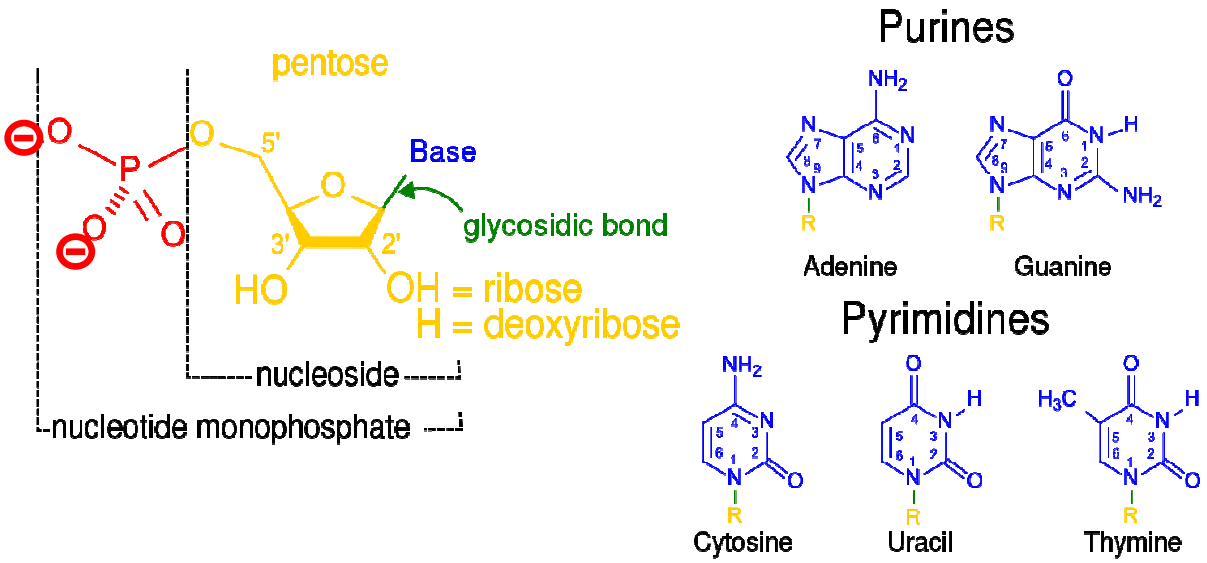


Figure 2.1: Structures of a nucleoside, nucleotide and the bases pyrimidine and purine found in DNA and RNA [12]

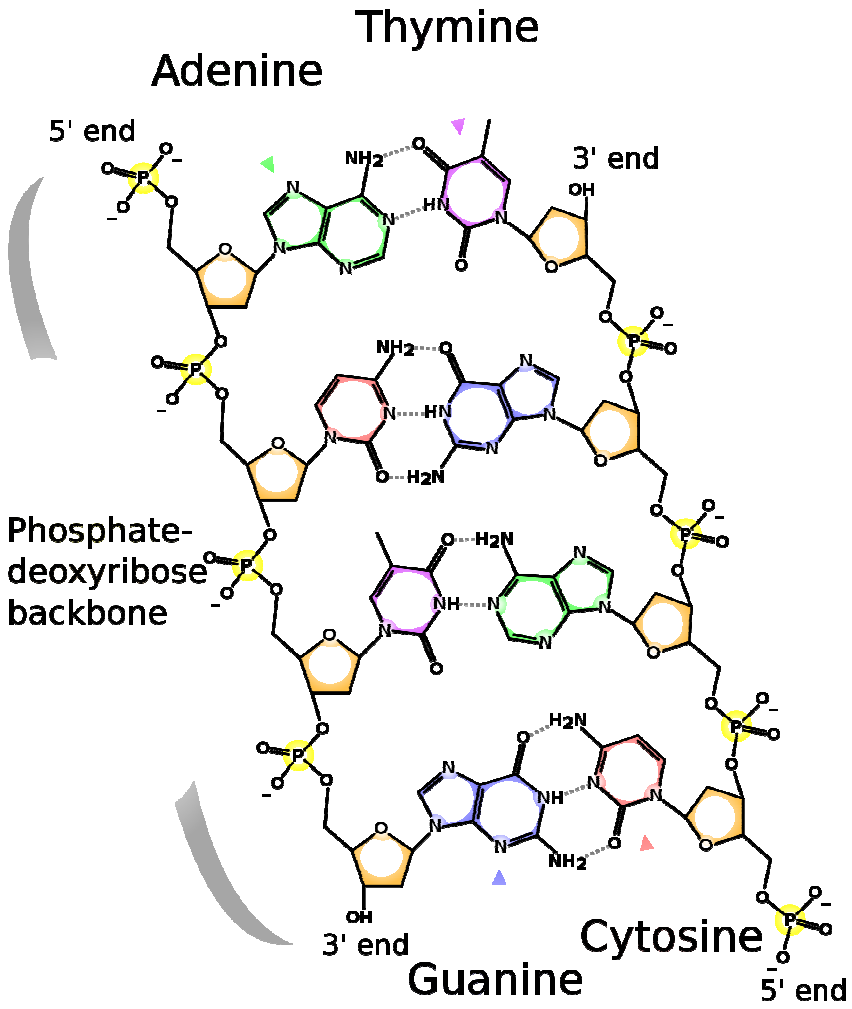


Figure 2.2.: Scheme of the base pairing in DNA. The complementary bases form hydrogen bonds. A always base pairs with T and C always with G. The two polynucleotides are oriented anti-parallel, one in 3' to 5' direction, the copmplementary strand is oriented in 5' to 3' direction [12].

3 Capillary Electrophoresis (CE)

3.1 Principles

Electrophoresis is the migration of charged species under the influence of an electric field. The migration can be carried out either in a capillary or a non-convective surrounding. Due to the differing migration velocities the analytes are separated in the electrical field. Figure 3.1 illustrates the basic instrumental set-up of a CE system. The separation of the analytes takes place inside a thin fused silica capillary (25-100µm I.D.). Depending on the analyte to be investigated, the inner surface of the capillary may be coated or uncoated. The coating influences the quantity and the direction of the electroosmotic flow (EOF). In the case of an uncoated fused silica capillary the silanol groups on the inner surface lead to the formation of an EOF, depending on their degree of dissociation. Both ends of the capillary tube, which is entirely filled with the backgroundelectrolyte (BGE), are placed into the BGE reservoirs. Via the two electrodes a potential difference of up to 30 kV is applied from end to end of the capillary. As a consequence, an electrical field is built up which induces an electric current. Sample injection (a few nL) is done by removing the buffer reservoir at the inlet and placing the inlet of the capillary into the sample vial. The sample is introduced into the capillary by a height difference of the inlet and outlet vial, by applying pressure to the inlet vial, by vacuum at the outlet vial, or by electrokinetic injection which uses an electrical field.

Since the sample ions have different mobilities, they migrate through the capillary at different velocities. Consequently, the ions are separated in their relative migration through the capillary, if their differences in charge to mass ratio are distinct enough. The net migration velocity of a sample ion is the sum of its electrophoretic velocity plus the electroosmotic velocity of the solution. The electroosmotic flow (EOF) is equal for all analytes and therefore does not contribute to the separation process itself, *i.e.* neutral species migrate with the velocity of the EOF. Depending on degree of dissociation and on the inner surface of the capillary wall the EOF can vary in size and direction but is equal for all sample ions.

There are various methods for the detection of the sample ions. Very frequently UV absorbance detectors or laser induced fluorescence (LIF) detectors are used. In the case of UV detectors a light beam is used to measure the absorbance of the sample ions that pass by the detection window. In a LIF detector an excitation beam induces fluorescence in the sample. The emission of the sample is measured at a right angle to the excitation source. These

detectors allow an on-line detection of the samples, since they measure the samples as they pass by the detection window.

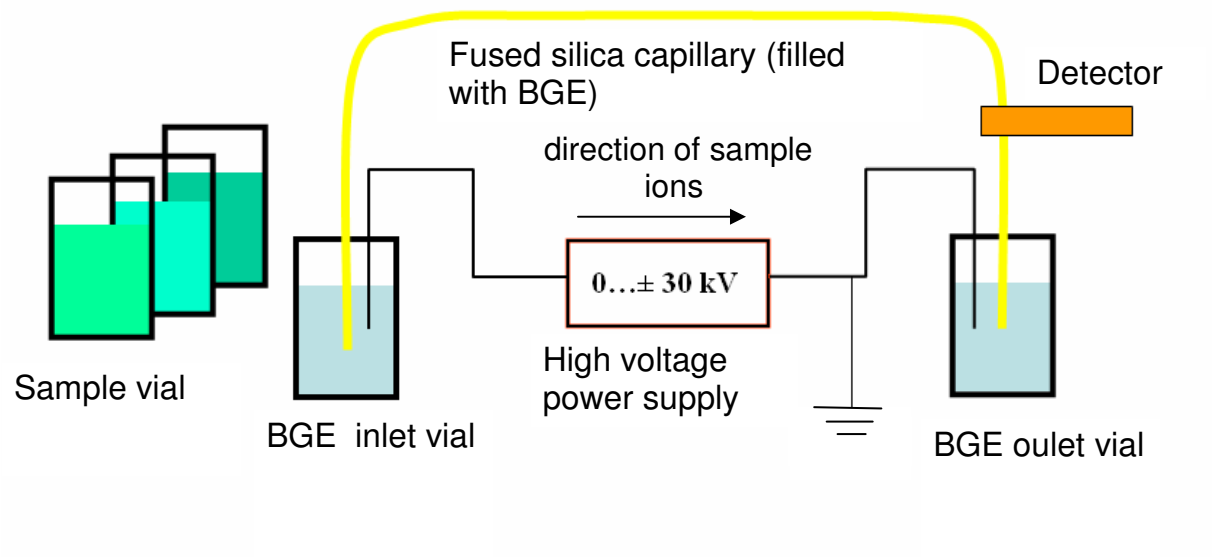


Figure 3.1: Instrumental set-up of a CE system

3.3 Theory of Electrophoretic Migration

A charged component, which is placed into an electric field, is accelerated by the electric force F_e that is proportional to the electric field strength.

$$F_e = \frac{z \cdot F \cdot E}{N_A}$$

F_e	electric force
z	charge number
F	Faraday constant
N_A	Avogadro constant
E	electric field strength

The electric force is counterbalanced by the friction force F_d , which is determined by Stoke's law.

$$F_f = 6 \cdot \pi \cdot \eta \cdot r \cdot u_{ep}$$

F_f	friction force
η	dynamic viscosity
r	Stoke's radius
u_{ep}	electrophoretic migration velocity

After a short time equilibrium between the electric force and the friction force is reached.

$$F_e = F_f$$

If we transform this equation we obtain:

$$u_{ep} = \frac{z \cdot F \cdot E}{N_A \cdot 6 \cdot \pi \cdot \eta \cdot r} = \frac{z \cdot e \cdot E}{6 \cdot \pi \cdot \eta \cdot r}$$

Since $\mu_{ep} = \frac{u_{ep}}{E}$ the electrophoretic mobility μ_{ep} can be calculated by:

$$\mu_{ep} = \frac{u_{ep}}{E} = \frac{z \cdot e}{6 \cdot \pi \cdot \eta \cdot r}$$

u_{ep} electrophoretic migration velocity
μ_{ep} electrophoretic mobility
e elemental charge

Depending on the surface of the inner capillary wall a certain EOF, which is superimposed on the electrophoretic migration, will arise. Electroosmosis causes the flow of the entire BGE, depending on the electric field strength. The EOF is caused by the surface charge on the inner capillary wall. In case of a fused silica capillary the negative surface charge is caused by deprotonated silanol groups. The BGE cations balance the negative surface charge and thereby form a mobile layer with opposite charge (Figure 3.2).
When an electric field is applied a flow of the entire mobile layer towards the cathode is created. The EOF can be described by the Helmholtz equation. Its quantity can be measured by measuring the migration velocity of a neutral substance, which is called an EOF marker. The net velocity u_{app} is the vectorial sum of u_{ep} and u_{eo} (Figure 3.3) and it is this that describes the separation of the ions in the analyte sample solution.

$$u_{eo} = \frac{\varepsilon \cdot E \cdot \zeta}{4 \cdot \pi \cdot \eta}$$

ε dielectric constant
η dynamic viscosity
ζ zeta potential
u_{eo} electroosmotic velocity

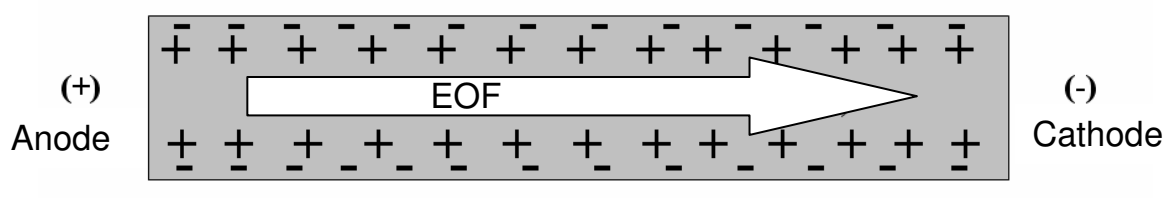


Figure 3.2: Formation of the EOF in a CE system

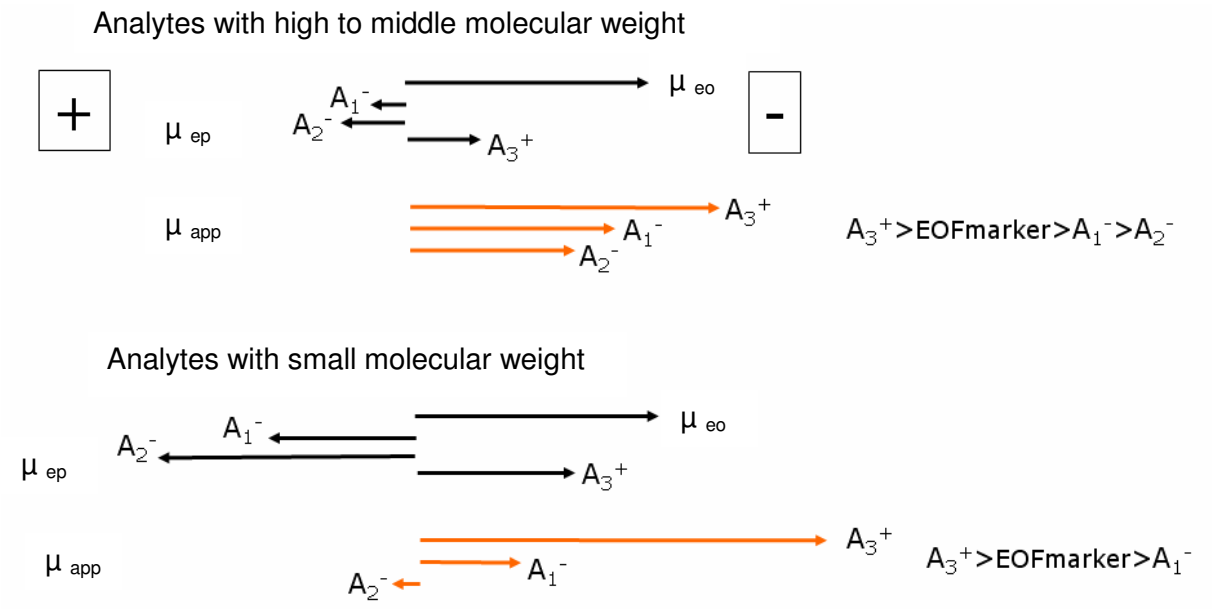


Figure 3.3: Apparent mobility as the vectorial sum of u_{eo} and u_{ep}

3.4 Electrophoretic Migration of DNA

The separation of polyelectrolytes has to be done according their differences in size. In a free solution of electrolyte no separation will occur. In the case of DNA, this is due its primary structure. DNA possesses a negatively charged phosphate deoxyribose backbone and therefore, the friction force F_f and the electrical force F_e increase linearly with the molecular mass M [9] [11].

To overcome this problem there are two methods. One concept is to dissolve molecular mechanical obstacles in the BGE which act as a sieving media that allows a size and conformation dependent separation, much like conventional agarose slab mediated electrophoresis that is routinely employed in molecular biology laboratories today. The sieving media can consist of physical or mechanical gels as well as solutions of linear polymers. In another approach, the concept of the end-labelled DNA fragments separated in free solution, an uncharged polymer is attached to the polynucleotide molecule. Consequently, the constant ratio of the charge to the friction is changed. This results in different electrophoretic mobilities [13].

The separation mechanisms of DNA in sieving media can be described by three migration regimes (see Figure 3.4) [9]. When radius of the hydrodynamic equivalent sphere, R_s, of the DNA polymer coil is smaller than the average mesh size, m, of the network the migration is controlled by accidental interactions between the DNA and the sieving medium obstacles (see Figure 3.4 A). If the DNA molecule is too large to fit inside a single pore (R_s is in the range of m) it is forced to deform and to reptate head first through the sieving media [11]. The term reptation refers to the polymer's snake like motion along a fictitious tube threaded through an array of fixed obstacles [14]. This means, it is either elongated or it expands in the void regions. The migration velocitiy in this regime strongly depends on the length of the DNA chain (see Figure 3.4 B). When R_s exceeds m multiple times the DNA polymers orient in field direction and migrates head first through the maze nearly all the time (see Figure 3.4 C). A separation in this area is nearly impossible, due to the fact that the sieving effect of the matrix is no longer active.

The three regimes can be described by the dependence of the logarithm of the relative electrophoretic mobility, $\frac{\mu}{\mu_0}$, on the logarithm of the molecular mass M (see Figure 3.5). In regime A, called the Ogston migration regime [15], the mobility of the DNA molecule is inversely proportional to the probability of collision with the obstacles of the sieving network.

The DNA molecule is not getting "hooked" in the matrix in this regime; it collides with the network only one place at a time. Regime B is the area with the strongest dependence of the mobility on *m*. As a result the separation selectivity is the highest in this regime. At higher molecular mass the migration of the DNA polymer underlies the so-called biased reptation with stretching [16]. The mobility tends to be size independent since a permanent stretching of the molecule occurs, which leads to a decrease in separation selectivity (see Figure 3.4 C).

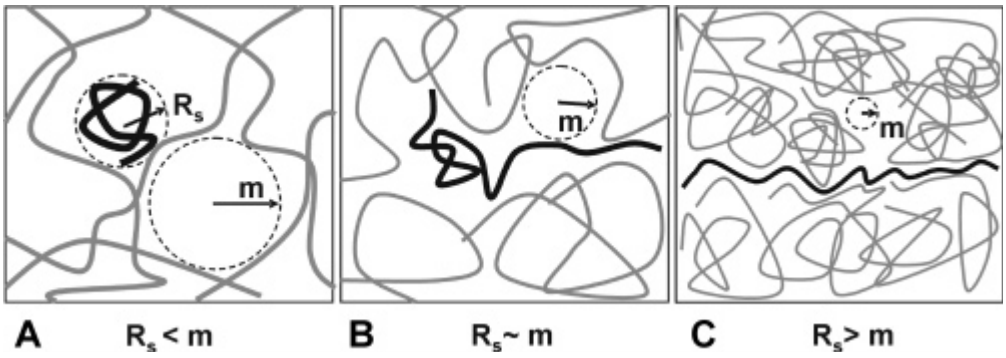


Figure 3.4: Schematic interpretation of the migration behaviour of DNA molecules depending on its R_s to m proportion in the migration regimes A, B and C [9]

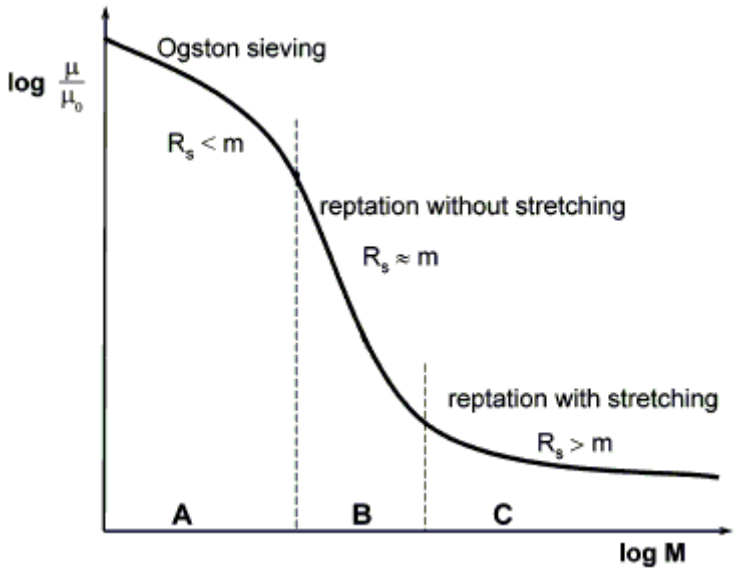


Figure 3.5: Dependency of the logarithm of the relative electrophoretic mobility, $\frac{\mu}{\mu_0}$, on the logarithm of the molecular mass M and the three different migration regimes A,B and C [9]

3.4.1 Sieving Matrices for DNA Separations

Sieving media have a vast influence on the performance of DNA separations [11]. They allow a separation of polyelectrolytes, such as DNA, according their differences in size, as long as the proportion of R_s to m is not as high as in regime C (see Figure 3.5). They can consist of physical or chemical gels as well as solutions of linear polymers.

Chemical and physical gels were the first matrices that were used for the DNA separation in capillaries [17]. Among the group of the chemical gels there are gels that consist of cross-linked polyacrylamide. Possibilities to vary the physical properties of the gel are the use of different cross-linkers or various derivatives of acrylamide monomers. In general acrylamide matrices are sieving media with small pores sizes which range from a few nanometres in concentrated gels to up to tens of nanometres in diluted gels [18]. Due to their small pores they are suitable for the separation of oligonucleotides. Agarose gels belong to the group of the physical gels. The closing of the pores is caused by the entanglement of the polysaccharide helices, whereas in chemical gels, such as cross-linked acrylamide, the cross linking agent reacts chemically with the polymer chains to vary pore size.

When using gel-filled capillaries several aspects have to be considered. Once the gel is polymerized inside the capillary it can not be replaced between the analyses. This leads to a decrease in reproducibility of the measurements. This is due to changes in the chemical or physical structures, alkaline or acidic hydrolysis, mechanical damage in flexible capillaries, or by deterioration of the gel by impurities or bubble formation as a result of Joule heating. When an electric current passes through a conductor, heat is released. This process is termed Joule heating. Another drawback of gel filled capillaries is the fact that the overall conductivity of the capillaries increase over time due to ion depletion at the capillary ends. Therefore gel capillaries are used only for three or four analyses [9]. In some cases they are even only used for a single run, in order to reduce reproducibility problems.

The employment of polymer solutions solved the common drawbacks of gel filled capillaries. The advantages of these true polymer solutions are the following: they can be easily replaced after each run and thus provide identical analytical conditions each run; mechanical destruction of the sieving matrix in flexible capillaries is avoided; it is possible to introduce the sample not only by applying a voltage, as in the case of gel filled capillaries, but also hydrodynamically; when non-coated capillaries are used the EOF can be used to change the migration order; the type of polymer and its concentration can be altered in consecutive runs; automation is facilitated by the use of polymer solutions; the selectivity of separation runs carried out in capillaries filled with polymer solutions is just as high as in gel capillaries [9].

As polymers, liquefied agarose [19] and other polysaccharides, linear polyacrylamide [20], poly(ethylene oxide) [21], poly(vinyl pyrrolydine) [22], and various alternatives to polymer solutions can be used.

Overall, in a comparison of conventional automated slab gel electrophoresis with gel-filled capillaries, polyacrylamide gel-filled capillaries yielded better resolution besides and three time's faster migration times [23]. To overcome the low stability and limited lifetime of gel-filled capillaries polymer solutions can be used, which allow high throughput DNA separations, such as those required in DNA sequencing technologies (see below). Research in this area revealed that highly entangled solutions of hydrophilic, high molecular mass polymers allow high separation efficiencies [24].

3.4.2 Fluorescent Labelling

The intrinsic fluorescence of nucleotides has a relatively low quantum yield. Consequently, covalently attached fluorescent tags or noncovalent staining dyes are used to label DNA samples in CE analyses. These labels provide a sensitive marker that can then be readily detected in the CE apparatus.

Covalently attached labels have to fulfil several criteria: maximum possible distance between the absorption and emission maxima of the fluorophor to ensure low background due to excitation source light scattering; the emission maxima of any different labels within the same sample should be as far from each other as possible, in order to allow clear identification; high quantum yields are best to provide good detection sensitivity; the annealing of primers or the incorporation of terminating dideoxynucleoside triphosphates should not be affected by the conjugation of a fluorophor; the labels should not significantly influence the electrophoretic mobility of the DNA fragments. Two of the most frequently used labels are fluorescin and rhodamine that both have emission wavelength within the visible range. Via a nucleophile addition of the DNA's amine group to the isothiocyanate group of the dye a covalent linkage is formed. The structures of fluorescin and rhodamine with their reactive isothiocyanate groups, and the reactive amine group of the DNA fragments, such as primers or 2',3'-dideoxynucleotides, are illustrated Figure 3.6.

Non-covalent dyes interact with DNA via complexation and are usually inserted between the neighbouring bases in a double stranded (ds) or single stranded (ss) DNA chains. As a result the conformation changes and the migration time and the separation selectivity increases. They are usually intercalated between every 4-10th bp in dsDNA and this leads to an increase in fluorescence signal with molecular size. DNA conformation can also play an important role

as a slightly changed conformation can result in a greater than 1000-fold fluorescence enhancement [25]. Another important property of non-covalently attached labels is the fact that they are unselective and will bind to any nucleic acid species. They typically possess a planar monomeric or homodimeric structure. Ethidium bromide is the most commonly used dye in CE, its structure is illustrated in Figure 3.7.

Several other strategies which use near infrared fluorescence detection [26], time resolved fluorescence [27], the implementation of resonance fluorescence energy transfer [28], have been described in the literature.

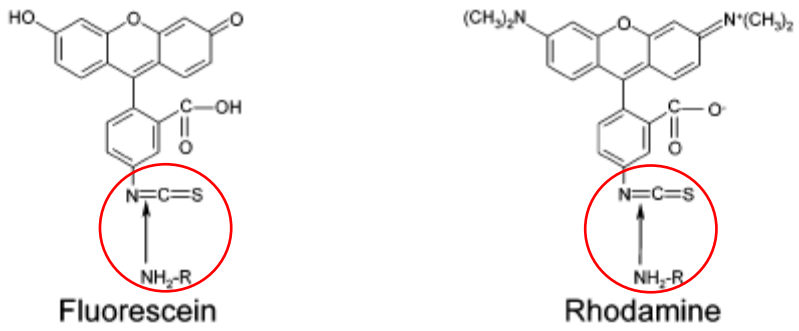


Figure 3.6: Structures of fluorescin and rhodamine and their reactive groups (in the highlighted areas) for the covalent linkage to the amine groups of primers or dideoxynucleoside triphosphates [9]

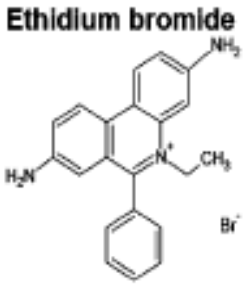


Figure 3.7: Structure of the planar monomeric intercalating fluorescent dye ethidium bromide [9]

3.4.3 Detection Systems

Although there are several methods, based on different principles, which are used for the detection of labelled DNA fragments in CE, Laser induced fluorescence (LIF) has emerged as the method of choice in medical diagnostics [29]. Although, there are also electrochemical [30] and UV absorbance [31] detection systems as well as mass spectrometry (MS) based

detectors connected to CE systems [32]. The simplest arrangement of an LIF detector consists of a laser on column excitation where the fluorescence emission is collected at a right angle. Further there exist confocal detectors and sheath flow detectors [9]. In a confocal detection system a laser beam, which is reflected by a dichroic mirror, is focused on the capillary by a microscopy objective and excites the samples that passes by the detection window. The fluorescence emission is collected by the same microscopy objective and passes the dichroic mirror. The fluorescence emission has to be detected without the excitation light. The elimination of the excitation light can be achieved by the the principle of confocal detection, which is illustrated in Figure 3.8. With pinhole optics (aperture) the excitation light, which is out-of-focus, and the scattered laser light are eliminated.

In the sheath flow detection system the detection spot is positioned behind the capillary outlet, in order to eliminate laser scattering from the capillary walls. The analyte is transported downstream by the flow along the capillary. Then the sample analytes are detected beneath the capillary tip in a wall less cell. The scheme is illustrated in Figure 3.9. The lower tubes provide electric contact as well as proper hydrodynamic flow of buffer. The fluorescence is the detected at right angles.

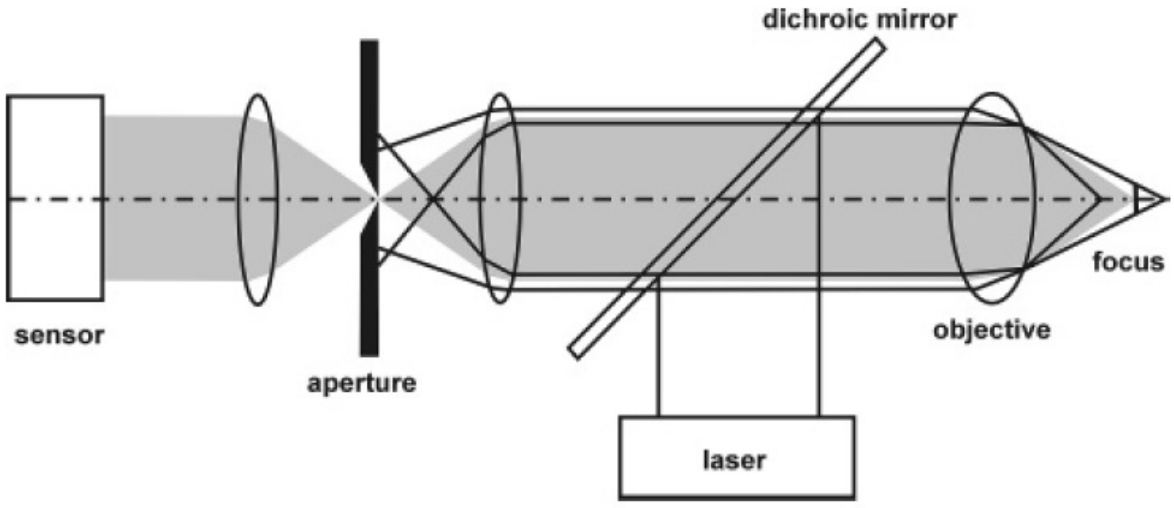


Figure 3.8: Scheme of a confocal detection system [9]

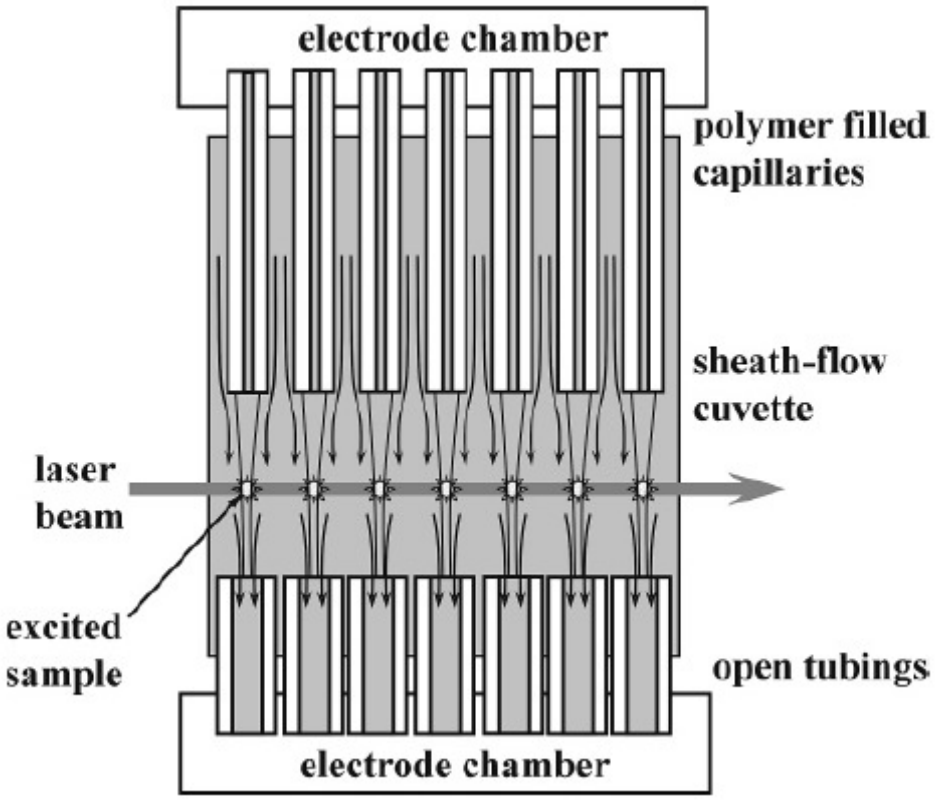


Figure 3.9: Scheme of sheath-flow detection system [9]

4 CE Methodology for Genome Analysis

4.1 DNA Sequencing

The goal of sequencing is to determine the precise sequence of bases that make up a DNA molecule. The knowledge of this sequence is important if we want to understand how different gene products influence the activity of each other within the whole organism. With only four different nucleotides, the nature of the DNA is relatively uniform. Paradoxically, the lack of chemical complexity in DNA molecules makes them rather difficult to sequence by means of more traditional methods adopted in protein biochemistry e.g. MS of fragmented samples that can be later reassembled to the whole.

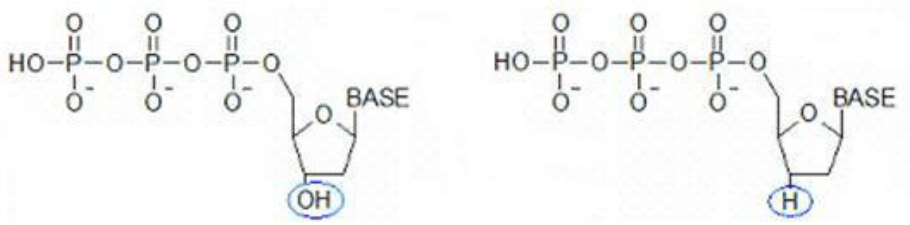
One method to determine the correct sequence is the Sanger sequencing reaction followed by CE separation of the fluorescently labelled reaction products. The theoretical principle of this method is that the order of bases/nucleotides is transformed into a code of specific lengths of DNA fragments that correspond to their sequence. These fragments can then be separated by CE according their differential sizes and this information used to decipher the underlying DNA sequence. A description of the experimental strategy is thus; The sample dsDNA that is to be sequenced is heated to 95°C, resulting in the formation of two ssDNA strands. In the next step, the target ssDNA strands are primed with a small 5' dye-labelled oligonucleotide sequence with a free 3' hydroxyl-group. This is achieved by reducing the temperature to 50-60°C; thus allowing the 5' dye-labelled oligonucleotide primer to anneal, via specific base pairing hydrogen bonds, to the template strand. A thermo-stable DNA polymerase then forms a complex between the template and the free 3' hydroxyl-group end of the primer and when the temperature is raised to approximately 70°C it starts to replicate the template strand by incorporating 2'-deoxynucleotides, using the ssDNA target strand as a template. However, the reaction mixture also contains 2',3'-dideoxynucleotides in addition to the normal 2'-deoxynucleotide substrates (the structures of deoxynucleotides and their 2',3'-dideoxyderivatives are illustrated in Figure 4.1). This ensures that when a 2',3'-dideoxynucleotide is incorporated, rather than the normal 2'-deoxynucleotide, into the new DNA strand the replication is terminated. This is because the incorporated 2',3'-dideoxynucleotide lacks the necessary free 3' hydroxyl-group required by the DNA polymerase to catalyse the addition of the next nucleotide. Hence the DNA synthesis ends at this specific 2',3'-dideoxynucleotide base incorporation. As 2',3'-dideoxynucleotides are randomly incorporated throughout the synthesis, the final reaction products are represented by newly replicated template DNA fragments of varying lengths, that are randomly terminated

by incorporation of 2',3'-dideoxynucleotides. Performing this experimental scheme four independent times for each DNA sample to be sequenced, in which each of the four replicates uses a reaction mixture containing only one of each of the four forms of 2',3'-dideoxynucleotides (*i.e.* A, G, T and C) results in a library of DNA fragments of differing lengths. Encoded within, this library it the underlying DNA sequence information. The newly replicated fragments of each reaction mixture, that contain only one of each of the four forms 2',3'-dideoxynucleotides (*i.e.* A, G, T and C), are fluorescently labelled. These fragments are then injected as a sample into a CE separation system. They are then separated according their differing lengths and detected via fluorescence detection. By comparison of the migration times of the products, from each of the four reactions, in four different CE runs (one for each 2',3'-dideoxynucleotide used), one can easily determine the sequence of the original sample/ template DNA (see Figure 4.2. A).

One way to increase the throughput of DNA sequencing is to use different terminator nucleotide concentrations. Four separate reactions, each using a different kind of terminator but the same 5' dye-labelled primer are performed. The important aspect is that the 2', 3'dideoxynucleotides concentrations in the individual reactions differ. This allows the four reaction products to be separated in a single CE run. Since the quantities of the reactions differ, the peak heights in the electropherogram can be used to identify the individual bases (see Figure 4.2 B).

Another approach is to label the 3' oligonucleotide primers with four different fluorescent dyes. Four separate reactions, each using a different primer and a different kind of 2',3'-dideoxynucleotide, are performed. Then the products can be pooled ,separated and detected in a single electrophoretic run by CE. This is because the differently labelled primers serve as markers for the individual bases (see Figure 4.2 C).

Most sequencing today is done by a method where four differently fluorescently labelled 2',3'-dideoxynucleotide 'terminators' are used. This approach has the advantage that only one sequencing reaction/ DNA polymerisation reaction is necessary, that can then be separated in a single CE run (see Figure 4.2 D). Another advantage of this method is that strands resulting from 'non-terminator' mediated mechanisms *e.g.* sudden stops due to polymerase pausing or disassociation, are not labelled.



deoxynucleotide triphosphate 2', 3' dideoxynucleotide triphosphate

Figure 4.1: Structures of a deoxynucleotide triphosphate and a 2', 3' dideoxynucleotide triphosphate

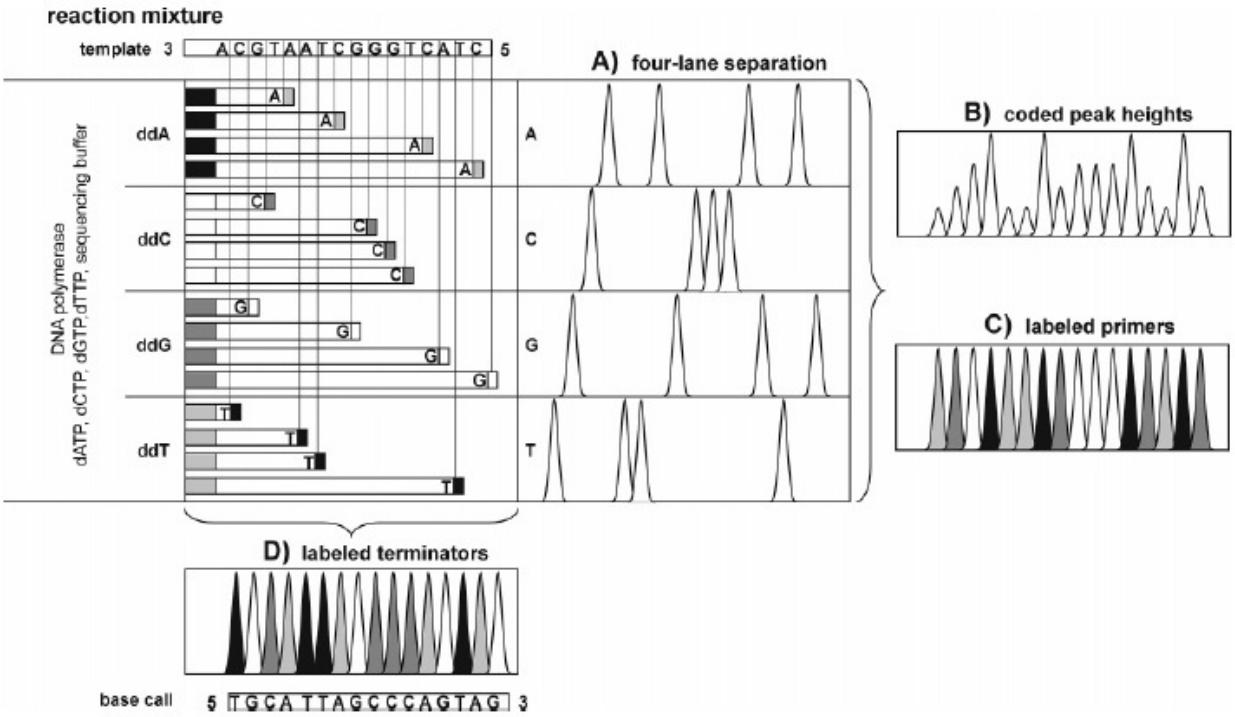


Figure 4.2: Scheme of four different separation and labelling strategies used in the Sanger sequencing reaction [9]

4.1.1 Application of DNA Sequencing

As described in the previous chapter, the results of CE analysis are many relatively short reads of DNA sequence of up to 600-700 bps. In order to perform DNA sequencing on a large scale it is necessary to find a good strategy for assembling those pieces.

One strategy to speed up large-scale DNA sequencing is the primer walking technique. For this a certain portion of the sequence has to be known, in order to be able to find a primer that will anneal to this known region. This primer is then used to sequence the adjacent regions. Based on the sequence of the adjacent regions a second primer is then synthesized. The replication step will then reveal the sequence beyond the known regions. This procedure is performed repeatedly. As a consequence the sequence of long DNA stretches can be resolved. One shortcoming of this method is that it is laborious and time consuming and requires the synthesis of many new and specific primers, which has to be done after each sequencing cycle.

Another method is the whole-genome shotgun sequencing strategy. Hereby the entire genome is fragmented either enzymatically or physically into short DNA fragments. These are then cloned into plasmids, sequenced, computationally analyzed for overlaps, aligned and assembled into a final sequence by computer software. The whole-genome shotgun sequencing strategy was applied in the Human Genome Project (HGP). This project was started in 1990 and its aim was to sequence the entire human genome within 15 years. It was finished under budget within only 10 years, owing to the use of capillary array electrophoresis (CAE) sequencers and the application of the whole-genome sequencing shotgun approach. CAE is a CE separation method that allows the parallel sequencing of many short DNA fragments in a multi capillary array. In the HGP 300 CAE sequencers (96-capillary ABI RPISM sequencer model 3700) were used parallel. This allowed the generation of a 14.8 billion bp DNA sequence from 27 271 853 high-quality sequence reads (550 nt) within 9 months in the year 2000 [33].

4.2 DNA Polymorphism Analysis

The detection of polymorphisms and mutations is of major importance for the characterization and diagnosis of human genetic diseases, and in many areas in molecular biology, genetics and medicine. There exists a variety of different techniques for mutation and polymorphism detection, where CE is centrally involved. The different approaches for the detection of mutations and polymorphisms described here are the restriction fragment length polymorphism (RFLP), the amplified fragment length polymorphism (AFLP) and the single strand conformation polymorphism (SSCP). Besides these strategies, there exist many other methods where CE can be applied: Examples are the denaturing gradient and the constant denaturant capillary electrophoresis [9] as well as the single nucleotide primer extension [10].

4.2.1 Restriction Fragment Length Polymorphism Analysis

One method in DNA diagnostics is to determine the polymorphism of DNA fragments. The test fragments are generated by digesting genomic DNA with restriction enzymes. The enzymatic digestion of the genomic DNA reveals a specific number of fragments with specific lengths. When the genomic DNA of different individuals, populations or species is split by the same restriction enzymes a different pattern of DNA fragments is obtained. This is due to the different positions of insertions, deletions and base substitutions in the genomic DNA. A deletion or insertion can lead to differences in the fragment lengths owing to the gain or loss of specific restriction sites in the genomic DNA sequence (see Figure 4.3). Although it should be noted that base substitutions are more difficult to detect because they do not cause changes in fragment sizes. However, they can cause the creation of new restriction sites, or if the point mutation happens at a potential restriction site, this site can be effectively removed as illustrated in Figure 4.4. CE is used to determine the length polymorphism of the fragments via their corresponding mobility differences. The information about the fragment length polymorphism can be used to either detect mutations or to construct physical maps of genomes [9].

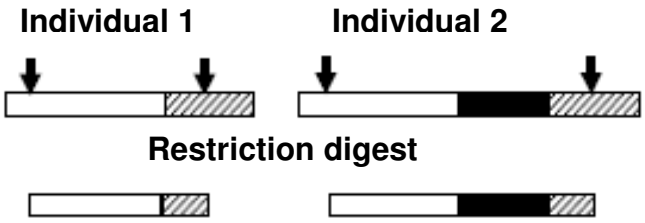


Figure 4.3: Scheme of the RFLP method. Gene sequences of two individuals treated with the same restriction endonuclease reveal fragments of different length due to an insertion [34].

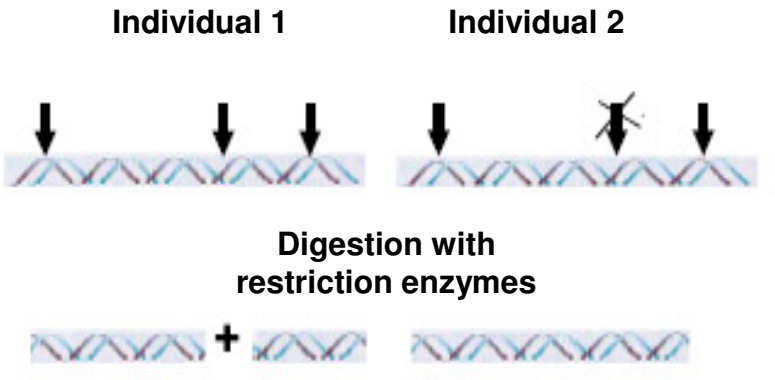


Figure 4.4: Scheme of the RFLP method. Gene sequences of two individuals treated with the same restriction endonuclease reveal different products due to the removal of a restriction site. The loss is caused by a point mutation [34].

4.2.2 Amplified Fragment Length Polymorphism Analysis

The amplified fragment length polymorphism is a PCR based method for the detection of the length polymorphism of DNA fragments. The scheme of this method is illustrated in Figure 4.5. First the genomic DNA is digested by restriction enzymes (very often MseI and EcoRI). Further oligonucleotide adaptors of known sequence, with sticky ends complimentary to the sticky ends of the enzymatically digested genomic DNA, as well as a ligase are added to the reaction mixture (see Figure 4.5 a). The ligase is an enzyme that catalyses the joining of the adaptors to the ends of the digested DNA fragments. The restriction and ligation step yields a huge amount of fragments with adaptors of known sequence at each end, but the sequences between the restriction sites are unknown (see Figure 4.5 b). In order to select a specific proportion of fragments, primers are annealed in the next step. Their sequence is complimentary to the following parts: the adaptor sequence; the remaining part of the restriction site; some bases inside the restriction site of the genomic DNA. Consequently, the primers will only anneal to fragments with the correct sequence. That means that primers determine the size of the fragments in AFLP. Then these selected fragments are amplified via PCR (Figure 4.5 c) and further analyzed be CE. CE has successfully been adapted to this standard analytical technique. The fragments of multiple DNA species are separated parallel according their size differences. Afterwards the results are compared in order to find polymorphisms in fragment profiles [10].

In AFLP polymorphism of the fragments occurs when: insertions or deletions occur either at the restriction site, between the restriction sites, or at the primer site; base substitutions appear at the restriction or primer site.

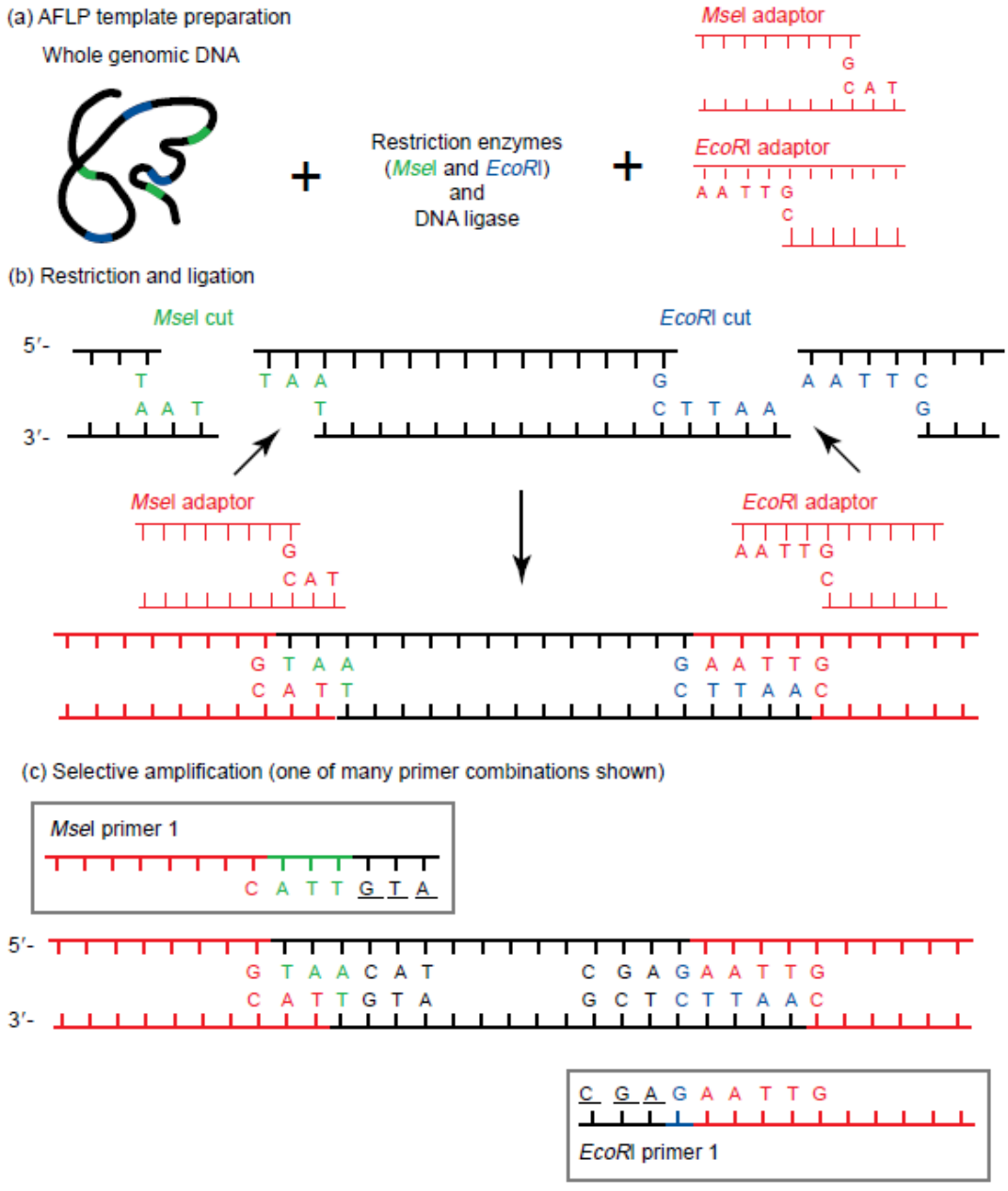


Figure 4.5: Scheme of the AFLP method. In (a), the genomic DNA is fragmented via restriction enzymes. In step (b) the adaptors are ligated. Finally, fragments, which have the correct sequence, are annealed to the primers and then amplified by PCR (c) [35].

4.2.3 Single Strand Conformation Polymorphism Analysis

Another method for the diagnostic screening of mutations and polymorphism is the single-strand conformation polymorphism (SSCP). This method relies on the principle, that single stranded fragments form a folded conformation determined the specific nucleotide sequence [36]. When the nucleotide variant affects the folding, the specific mobility of the single strand is altered.

First of all the DNA sample is denatured, then the denatured ssDNA fragments are introduced into a non-denaturing sieving medium and separated by CE. The non-denaturing sieving medium causes the ssDNA fragments to form a folded conformation, depending on the individual sequence, which leads to differences in the specific mobilities. However, if the mutation occurs in a single strand loop the mobility may not be affected. The principle is illustrated in Figure 4.6. The native dsDNA of the wild type and the point mutation have the same effective electrophoretic mobility, which is indicated by the arrows below the conformation structures. In a denatured environment the strands are separated and the ssDNA samples of the wild type and the mutant type have all equal effective electrophoretic mobilities. When those ssDNA samples are introduced into a non denaturing sieving matrice they undergo a conformational change. Thus the effective electrophoretic mobilities between the wild type and the mutant type differ.

The application of CE in this field allows fully automated analysis within a short time associated with high separation efficiency. However, the sensitivity of the mutation detection decreases with the length of the DNA fragments, since the conformation determines the mobility.

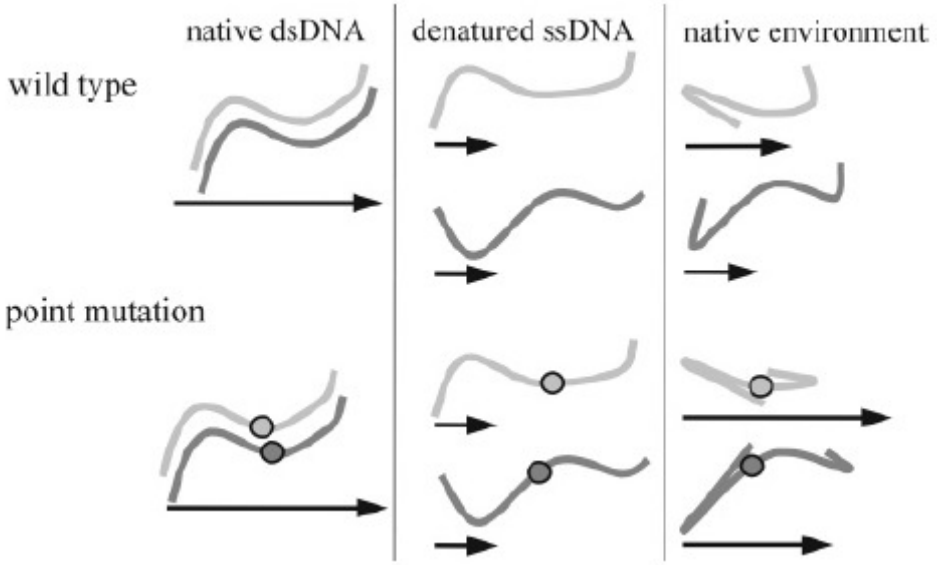


Figure 4.6: Scheme of the SSCP method. The conformations of the ssDNA fragments are sensitive to point mutations, this leads to a change in their mobilities [9].

5 Conclusion

The DNA-changes that a living organism accumulates during its live span can result in mutations and polymorphisms. These sequence variations can be used in forensic science and medicine to serve as an individual's barcode, to develop therapies on an individual's genetic profile, or in evolutionary biology. In this thesis various analysis techniques such as RFLP, AFLP, and SSCP for the detection of mutations and polymorphisms have been covered. CE offers a lot of possibilities in the area of mutation and polymorphism detection. Therefore it is used in AFLP, RFLP and SSCP analysis to detect the sequence variations. Sequencing of DNA is done by the Sanger sequencing reaction and subsequent analysis of the products by CE, in order to allow a determination of the sequence. The Sanger sequencing reaction followed by CE separation was applied in the HGP where the sequence of the human genome was sequenced in only 10 years. Within this project the human genome was fragmented randomly by the whole-genome shotgun method. The DNA sequences of the fragments were finally determined by using CE based DNA sequencers.

Overall it can be said, that the application of CE in the field of genetic analysis has led to the development of fully automated high performance CE devices that greatly benefit DNA sequencing and genotyping.

6 References

[1] Rampal, J. B. Ed. *DNA Arrays: Methods and Protocols*; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2001. 29.

[2] Darby, I. A., Hewitson, T. D., Eds. *In Situ Hybridization Protocols*; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2005.

[3] Devaney, J. M.; Girard, J. E.; Marino, M. A. *Anal. Chem.*, 72 (2000) 858.

[4] Banoub J. H., Newton R. P., Esmans E., Ewing D. F., Mackenzie G. *Chem. ReV.*, 2005, *105*, 1869.

[5] Hawley T. S., Hawley R. G. Eds. *Flow Cytometry Protocols*; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2004.

[6] DeFrancesco, L. *Anal. Chem.*, 2003, *75*, 175A.

[7] Ronaghi, M. *Genome Res.*, 2001, *11*, 3.

[8] Margulies, M.; Egholm, M.; Altman, W. E.; Attiya, S.; Bader, J. S.; Bemben, L. A.; Berka, J.; Braverman, M. S.; Chen, Y. J.; Chen, Z. T.; Dewell, S. B.; Du, L.; Fierro, J. M.; Gomes, X. V.; Godwin, B. C.; He, W.; Helgesen, S.; Ho, C. H.; Irzyk, G. P.; Jando, S. C.; Alenquer, M. L. I.; Jarvie, T. P.; Jirage, K. B.; Kim, J. B.; Knight, J. R.; Lanza, J. R.; Leamon, J. H.; Lefkowitz, S. M.; Lei, M.; Li, J.; Lohman, K. L.; Lu, H.; Makhijani, V. B.; McDade, K. E.; McKenna, M. P.; Myers, E. W.; Nickerson, E.; Nobile, J. R.; Plant, R.; Puc, B. P.; Ronan, M. T.; Roth, G. T.; Sarkis, G. J.; Simons, J. F.; Simpson, J. W.; Srinivasan, M.; Tartaro, K. R.; Tomasz, A.; Vogt, K. A.; Volkmer, G. A.; Wang, S. H.; Wang, Y.; Weiner, M. P.; Yu, P. G.; Begley, R. F.; Rothberg, J. M. *Nature*, 2005, *437*, 376.

[9] Kleparnik, K.; Bocek, P. *Chem. Rev.*, 2007, *107*, 5279-5317.

[10] Mitchelson, K.R. *Molecular Biotechnology*, 2003, *24*, 41-68.

[11] Slater, G. W.; Kenward, M.; McCormick, C. L.; Gauthier, G. M. *Current Opinion in Biotechnology*, 2003, *14*, 58-64.

[12] http://en.wikipedia.org/wiki/DNA, author: Madeleine Price Ball, (dl: 20.12.2010).

[13] Defontaines, A. D.; Viovy, *J. L. Electrophoresis*, 1993, *14*, 8.

[14] De Gennes, P.G. *J. Chem. Phys.*, 1971, *55*, 572 – 579.

[15] Ogston, A. G. *Trans Faraday Soc.*, 1958, *54*, 1754.

[16] Lumpkin, O. J.; Dejardin, P.; Zimm, B. H. *Biopolymers,* 1985, *24*, 1573.

[17] Cohen, A.S.; Najarian, D.; Smith, J. A.; Karger B. L. *J. Chromatogr.*, 1988, *458*, 323.

[18] Righetti, P. G.; Caglio, S.; Saracchi, M.; Quaroni. S. *Electrophoresis*, 1992, *13*, 587.

[19] Kleemiss, M. H.; Gilges, M.; Schomburg, G. *Electrophoresis*, 1993, *14*, 515.

[20] Sudor, J.; Foret, F.; Bocek, P. *Electrophoresis*, 199, *12*, 1056.

[21] Chang, H. T.; Yeung, E. S. *J. Chromatogr., B: Biomed. Appl*., 1995, *669*, 113.

[22] Pang, H.; Pavski, V.; Yeung, E. S. *J. Biochem. Biophys. Methods* 1999, *41*, 121.

[23] Spencer, M. *Electrophoresis* 1983,14.

[24] Wu, C.; Quesada, M. A.; Schneider, D. K.; Farinato, R.; Studier, F. W.; Chu, B. *Electrophoresis* 1996, *17*, 1103.

[25] Rye, H. S.; Yue, S.; Wemmer, D. E.; Quesada, M. A.; Haugland, R. P.; Mathies, R. A.; Glazer, A. N. *Nucleic Acids Res.* 1992, *20*, 2803.

[26] Lee, Y.H.; Maus, R. G.; Smith, B. W.; D., W. J. *Anal. Chem.* 1994, *66*, 4142.

[27] Waddell, E.; Lassiter, S.; Owens, C. V.; Soper, S. A. *J. Liq. Chromatogr. Relat. Technol.* 2000, *23*, 1139.

[28] Sapsford, K. E.; Berti, L.; Medintz, I. L. *Miner*V*a Biotecnol.* 2004,*16*, 247.

[29] Zhang, J. Z.; Chen, D. Y.; Wu, S.; Harke, H. R.; Dovichi, N. J. *Clin. Chem.* 1991, *37*, 1492.

[30] Brazill, S. A.; Kim, P. H.; Kuhr, W. G. *Anal. Chem.* 2001, *73*, 4882.

[31] Froim, D.; Hopkins, C. E.; Belenky, A.; Cohen, A. S. *Nucleic Acids Res.* 1997, *25*, 4219.

[32] Willems, A. V.; Deforce, D. L.; van Peteghem, C. H.; van Bocxlaer, J. F. *Electrophoresis* 2005, *26*, 1221.

[33] Venter, C. J.; Adams, M. D.; Myers, E. W.; Li, P. W.; Mural, R. J.; Sutton, G. G.; Smith, H. O.; Yandell, M.; Evans, C. A.; Holt, R. A.; Gocayne, J. D.; Amanatides, P.; Ballew, R. M.; Huson, D. H.; Wortman, J. R.; Zhang, Q.; Kodira, C. D.; Zheng, X. H.; Chen, L.; Skupski, M.; Subramanian, G.; Thomas, P. D.; Zhang, J.; Miklos, G. L. G.; Nelson, C.; Broder, S.; Clark, A. G.; Nadeau, J.; McKusick, V. A.; Zinder, N.; Levine, A. J.; Roberts, R. J.; Simon, M.; Slayman, C.; Hunkapiller, M.; Bolanos, R.; Delcher, A.; Dew, I.; Fasulo, D.; Flanigan, M.; Florea, L.; Halpern, A.; Hannenhalli, S.; Kravitz, S.; Levy, S.; Mobarry, C.; Reinert, K.; Remington, K.; Abu-Threideh, J.; Beasley, E.; Biddick, K.; Bonazzi, V.; Brandon, R.; Cargill, M.; Chandramouliswaran, I.; Charlab, R.; Chaturvedi, K.; Deng, Z.; Francesco, V. D.; Dunn, P.; Eilbeck, K.; Evangelista, C.; Gabrielian, A. E.; Gan, W.; Ge, W.; Gong, F.; Gu, Z.; Ping, Guan, T. J. H.; Higgins, M. E.; Ji, R.-R., Ke, Z.; Ketchum, K.

A.; Lai, Z. Y. L.; Li, Z.; Li, J.; Liang, Y.; Lin, X.; Lu, F.; Merkulov, G. V. N. M.; Moore, H. M.; Naik, A. K.; Narayan, V. A. B. N.; Nusskern, D.; Rusch, D.; Salzberg, S.; Shao,W. B. S.; Sun, J.; Wang, Z. Y.; Wang, A.; Wang, X.; Wang, J.; Wei, M.-H. R. W.; Xiao, C.; Yan, C.; Yao, A.; Ye, J.; Zhan, M.; Zhang, W. H. Z.; Zhao, Q.; Zheng, L.; Zhong, F.; Zhong, W.; Zhu, S. C. S. Z.; Gilbert, D.; Baumhueter, S.; Spier, G.; Carter, C. A. C.; Woodage, T.; Ali, F.; An, H.; Awe, A.; Baldwin, D. H. B.; Barnstead, M.; Barrow, I.; Beeson, K.; Busam, D.; Carver, A. A. C.; Cheng, M. L.; Curry, L.; Danaher, S.; Davenport, L.; Desilets, R. S. D.; Dodson, K.; Doup, L.; Ferriera, S.; Garg, N.; Gluecksmann, A. B. H.; Haynes, J.; Haynes, C.; Heiner, C.; Hladun, S.; Hostin, D. J. H.; Howland, T.; Ibegwam, C.; Johnson, J.; Kalush, F. L. K.; Koduru, S.; Love, A.; Mann, F.; May, D.; McCawley, S. T. M.; McMullen, I.; Moy, M.; Moy, L.; Murphy, B.; Nelson, K. C. P.; Pratts, E.; Puri, V.; Qureshi, H.; Reardon, M.; Rodriguez, R. Y.- H. R.; Romblad, D.; Ruhfel, B.; Scott, R.; Sitter, C.; Smallwood, M. E. S.; Strong, R.; Suh, E.; Thomas, R.; Tint, N. N.; Tse, S. C. V.; Wang, G.; Wetter, J.; Williams, S.; Williams, M.; Windsor, S. E. W.-D.; Wolfe, K.; Zaveri, J.; Zaveri, K.; Abril, J. F. R. G.; Campbell, M. J.; Sjolander, K. V.; Karlak, B.; Kejariwal, A. H. M.; Lazareva, B.; Hatton, T.; Narechania, A.; Diemer, K.; Muruganujan, A. N. G.; Sato, S.; Bafna, V.; Istrail, S.; Lippert, R.; Schwartz, R. B. W.; Yooseph, S.; Allen, D.; Basu, A.; Baxendale, J.; Blick, L. M. C.; Carnes-Stine, J.; Caulk, P.; Chiang, Y.-H.; Coyne, M.; Dahlke, C. A. D. M.; Dombroski, M.; Donnelly, M.; Ely, D.; Esparham, S.; Fosler, C. H. G.; Glanowski, S.; Glasser, K.; Glodek, A.; Gorokhov, M.; Graham, K. B. G.; Harris, M.; Heil, J.; Henderson, S.; Hoover, J.; Jennings, D. C. J.; Jordan, J.; Kasha, J.; Kagan, L.; Kraft, C.;Levitsky, A. M. L.; Liu, X.; Lopez, J.; Ma, D.; Majoros, W.; McDaniel, J. S. M.; Newman, M.; Nguyen, T.; Nguyen, N.; Nodell, M.; Pan, S. J. P.; Peterson, M.; Rowe, W.; Sanders, R.; Scott, J. *Science* 2001, *291*, 1304.

[34] Z.J. Liua; J.F. Cordesb *Aquaculture* 2004, *238*, 1 –37.

[35] U. G. Mueller; .L. L. Wolfenbarger *TREE* 1999, *14*, 10.

[36] Kozlowski, P.; Krzyzosiak, W. J. *Nucleic Acids Res*. 2001, *29*, e71.