3. Introduction

3.1 General introduction

The complexity of any organism is ultimately determined by its genes - their mere number and composition as well as their interactions and regulation. To unravel the complex network of molecular processes that underlies an organism's physiology it is essential to identify all its existing genes, elucidate their functions and comprehend their interplay.

Although a great deal of genomes has already been fully sequenced[1], and more are to come over the next years, the complete genome sequence on its own is not sufficient to understand the biology of an organism. Even many of the potential coding sequences within these completed genomes are not readily identifiable by gene prediction programs. Furthermore, correct splice patterns and splice variants are even more difficult to predict by computer analysis.

In contrast, cDNA libraries ideally represent all expressed genes specific for a biological source from which they originate (i.e., a cell type, a tissue, or a developmental stage). They provide direct access to the final splice variants and therefore correct protein products and give valuable insight into the abundance levels of different mRNAs in the starting material. As a consequence, the analysis of cDNA libraries does not only complement genomic sequencing but represents an important step towards a functional interpretation of genomic data. To guarantee that low-abundance mRNAs or very rare transcripts (~ one molecule per cell) are represented in a cDNA library, often consisting of several cell types, it becomes necessary to construct and screen libraries containing up to several million independent cDNA clones (Sambrook et al., 1989).

3.2 Sequencing by hybridization

The most common way of analyzing cDNA sequences in a high throughput fashion is expressed sequence tag (EST) sequencing (Adams et al., 1995) which is based on the conventional enzymatic chain termination method developed by Sanger et al. (1977). ESTs which are "single-pass" cDNA sequences may serve to enhance gene discovery, help annotate genomic sequence, and provide cDNA resources for mapping and

[1] These comprise more than 1000 viruses and microbes as well as several prominent eukaryotes, such as soil nematode *Caenorhabditis elegans*, mustard family plant *Arabidopsis thaliana*, and fruit fly *Drosophila melanogaster* (refer to: National Center of Biotechnology Information (NCBI), Entrez Genome Section, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome).

downstream functional analysis (Marra et al., 1998). However, large-scale EST projects demand an appropriate infrastructure and hence are very costly. Another limiting factor is the availability of suitable normalized libraries.

An alternative concept of sequence deciphering was independently proposed by several research groups in the late 1980s (Bains et al., 1988, Lysov et al., 1988, Drmanac et al., 1989): sequencing by hybridization (SBH). It takes up one of the innate properties of DNA that is, each single strand interacts with its uniquely complementary sequence via Watson-Crick base pairing – a phenomenon termed “hybridization”.

According to the concept of SBH, any linear sequence can be regarded as an assembly of shorter overlapping fragments, i.e. oligonucleotides. Consequently, it is – theoretically – possible to ultimately deduce the complete sequence of a given target from independent hybridizations with all possible (4^n) oligonucleotide probes of a given length *n*. The necessary sequence reconstruction is to be done by various computational algorithms in accordance with the overlaps of hybridized oligonucleotide probes.

The probe length ought to be reasonably short to prevent the total number of all possible combinations from being impractically large, however, not too short for an unambiguous determination of a target sequence. Therefore, octamer probes were considered to be most suitable as for SBH of double-stranded DNA a full probe set would consist of 32,768 ($4^8/2$) oligonucleotides taking into account the redundancy of complementary sequences (e.g. 5`-ATGCATGC-3` and 5`-TACGTACG-3`).

Improved sequence reconstruction algorithms, however, allowed to further reduce the total number of probes required. Drmanac and co-workers (1993) showed that instead of 16,384 heptamers (4^7) only 3000 appeared to be sufficient for accurate sequencing of 1.1 kb DNA. Although octa- and heptamers have the advantage of a better mismatch discrimination in hybridizations (i.e. they are more specific), their DNA duplex stability is significantly decreased compared to longer oligonucleotides. A number of methods was proposed to overcome this drawback of shorter probes, such as the introduction of chemically modified nucleotides (Hoheisel et al., 1990), the use of universal bases (Parinov et al., 1996) or the use of 5´- and 3´-end degenerated bases (Drmanac et al., 1992).

The concept of SBH has been implemented in two different ways. Both approaches focus on arraying of either of the components of hybridization – target DNA or oligonucleotide probe. In the first one target DNA attached to a solid support is hybridized with a set of labeled synthetic oligonucleotides (Drmanac et al., 1996) whereas in the second one a complete set of oligonucleotide probes is arrayed on a surface and hybridized to an individually labeled target DNA (Khrapko et al., 1989; Southern et al., 1992). The latter approach is theoretically more informative since a single experiment is

supposed to provide full sequence information for a given DNA target. On the other hand, it also requires the manufacturing of absolutely flawless oligonucleotide arrays.

Initially, the concept of SBH promised to enable a significantly faster sequence determination than achievable by the conventional Sanger method. The barriers and technical limitations of that intriguing direct way of sequencing became more obvious later. First, it is almost unfeasible to achieve absolute hybridization specificity for short oligonucleotides hybridizing to complex target DNA. This holds in particular true for the second approach where multiple heterogeneous oligonucleotides are hybridized in parallel. Ambiguity is introduced by additive contribution of mismatched base pairs of an imperfect match. The event of hybridization is a complex chemical reaction consisting of various interactions between DNA nucleobases that contribute to duplex formation and its stability. These interactions are influenced by the number of perfectly matched (fully complementary) and mismatched pairings, the contribution of “dangling ends” (Williams et al., 1994) and nearest neighborhood parameters (Wetmur, 1991). In experiments, the efficient discrimination of mismatches can be at least improved by tuning experimental conditions such as adjusting probe concentration, time and temperature of washing etc. (Drmanac et al., 1990). Second, low complexity regions such as polyA stretches or tandem repeats frequently distributed in genomic DNA are prone to create branching points leading to errors in the process of sequence reconstruction. Computational simulations have shown that complete sets of octamers would only allow the determination of 80% of random DNA sequences of 200 bp length (Khrapko et al., 1989).

3.3 Oligonucleotide fingerprinting

3.3.1 The principle of oligonucleotide fingerprinting

A simpler and more modest application of the original SBH concept is the generation of partial sequence information using an incomplete set of oligonucleotide probes of a given length. It is commonly referred to as partial SBH and forms the basis of the oligonucleotide fingerprinting (OFP) approach. The latter has been conceptually developed more than a decade ago (Poustka et al., 1986) and has been continuously methodically refined ever since.

During the process of OFP vectors of hybridization signal intensities – so-called oligonucleotide fingerprints – are generated for each clone of an arrayed cDNA or genomic DNA library. This is accomplished through a series of up to 250 successive hybridizations of short, radioactively labeled oligonucleotides. An illustration of the principle underlying OFP is depicted in figure 3.1.

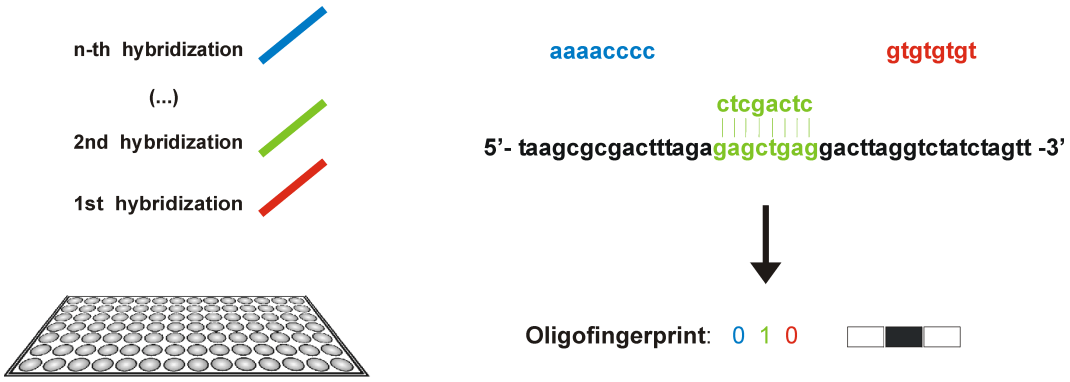

Figure 3.1 Principle of oligonucleotide fingerprinting

n radioactively labeled short oligonucleotides are successively hybridized to arrays of immobilized clones (PCR products). For each clone under investigation an oligofingerprint is generated reflecting the overall result of oligonucleotide hybridizations. Successful hybridization events are depicted as "1" or black boxes, unsuccessful ones are depicted as "0" or white boxes.

According to their individual fingerprints, clones are grouped into clusters. Ideally each cluster represents one gene with the size of the cluster indicating its relative level of expression (Drmanac et al., 1996, Meier-Ewert et al., 1998). For each cluster a consensus fingerprint is computed. The clone showing highest similarity to a consensus fingerprint is chosen as the representative clone for a cluster and sequenced. Via database comparisons ("in silico" fingerprinting) known genes are identified and potential candidate genes are discovered (fig. 3.2).

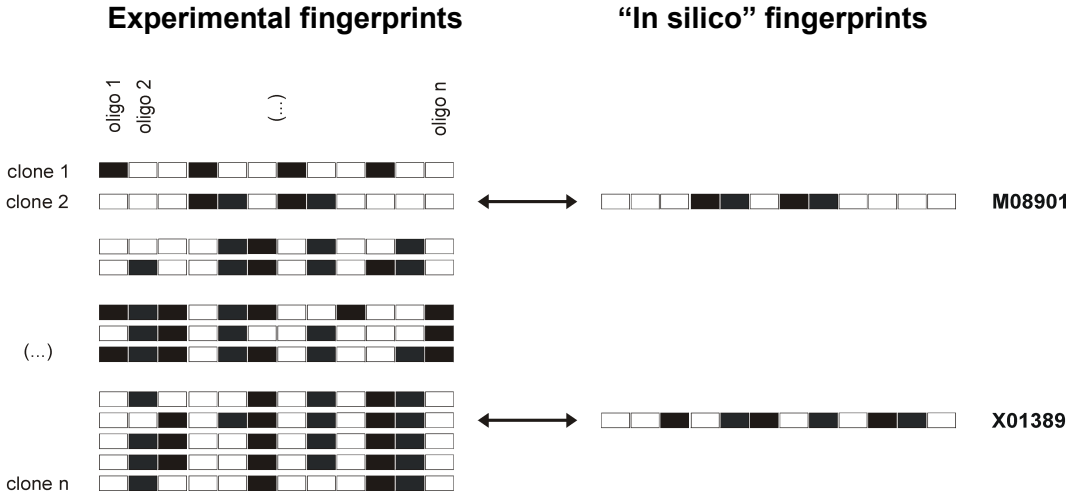

Figure 3.2 Outcome of oligonucleotide fingerprinting

Obtained fingerprints are subjected to a clustering analysis (Herwig et al., 1999) by which similar fingerprints are assigned to the same cluster whereas dissimilar ones are separated. Via comparisons with computationally generated "in silico" fingerprints of database sequences known genes are identified and potential candidate genes are discovered.

The complex process of OFP which consists of a crucial preceding hybridization probe design as well as several experimental and evaluation steps has been reviewed by Bauer et al. (2002). Comprehensive descriptions of the entire procedure are available at Meier-Ewert et al. (1998) and Clark et al. (1999).

3.3.2 Applications of oligonucleotide fingerprinting

OFP has been successfully employed in a wide range of applications, such as the identification of overlapping clones to construct ordered clone libraries (Craig et al., 1990) and the establishment of clone maps of large insert clones (Hoheisel et al., 1991).

A more recent application has been the reduction of redundancy experienced in large-scale shotgun sequencing projects (Radelof et al., 1998), where by an efficient combination of OFP and shotgun sequencing the number of shotgun clones to be sequenced is reduced at least 2-fold.

The suitability of OFP as a very effective means of differential cDNA library screening and powerful alternative to other differential expression technologies, such as EST analysis (Adams et al., 1995) and SAGE (Velculescu et al., 1995), has been demonstrated by Meier-Ewert et al. (1998) and Clark et al. (2001). Meier-Ewert and colleagues could identify several hundred clusters significantly differentially expressed in two cDNA libraries of two different stages of embryonic mouse development. Clark and co-workers, in their large comprehensive OFP study on an embryonic and an adult-liver zebrafish cDNA library, isolated and identified those transcripts specific to the respective library.

The most important application of OFP, however, is the normalization and characterization of cDNA libraries of a certain developmental stage, tissue or cell type of an organism under investigation. Thereby, non-redundant ”unigene” cDNA clone sets are derived, which represent the almost complete gene inventory of that organism or tissue thereof. For several reasons ”unigene” sets are highly valuable resources. They are well-defined and represent a non-redundant clone set which is directly accessible, e.g. in the form of bacterial clones. In addition to global sets, tissue-specific as well as development-specific subsets can be arrayed and used for various downstream applications, such as expression profiling using genome-wide or tissue-specific DNA microarrays as well as protein arrays. A variety of projects has taken advantage of the normalization effect achieved by OFP and the specific ”unigene” set which resulted therefrom. These projects comprise the normalization of embryonic mouse cDNA libraries (Meier-Ewert et al., 1998) and their subsequent gene expression profiling by complex hybridization (Eickhoff et al.,

2000), the use of OFP as a means of prescreening in the generation of a partial sea urchin gene catalogue (Poustka et al., 1999) and for a large-scale zebrafish EST project (Clark et al., 2001) as well as the construction of a ”unigene” set for sugar beet (Herwig et al., 2002), for which practically no molecular information had been available beforehand.

3.3.3 Assessment of normalization quality

Beside OFP other methods of library normalization exist. For small projects, a usual approach is to hybridize common cDNAs or a complex cDNA probe and then select the cDNAs that do not hybridize. This is a rapid technique, which should eliminate many of the most common cDNAs. However, it also selects for clones with no or small inserts and can eliminate related or repeat containing cDNAs (Adams et al., 1995, Gong et al., 1997).

For the most extensively used human cDNA libraries, the reassociation kinetics-based approach (Soares et al., 1994) has been successfully employed. However, it shows a tendency to select short, truncated, or internally primed inserts (Bonaldo et al., 1996), artifacts which can artificially inflate the apparent diversity of a library and number of genes identified (Ewing and Green, 2000), as well as to reduce the average insert size (Clark et al., 2001).

In their comprehensive OFP study on zebrafish cDNA libraries Clark et al. could also demonstrate that the normalization performance of OFP is superior to the one of the reassociation kinetics technique (fig. 3.3): more than two thirds of the generated ESTs of the normalized MPMGp609[2] library represent unique sequences. Furthermore, in recent OFP projects resulting EST sets contain almost 90% unique sequences (Herwig et al., 2002). This clearly illustrates the high normalization performance achieved by OFP.

3.3.4 Drawbacks of the current process

As a result of its multiple automated steps OFP technology can be readily employed in large-scale genomic applications. Still, the present technology shows some inherent limitations which need to be overcome in order to further increase the degree of automation and, hence, throughput.

Current bottlenecks are primarily the high number of oligonucleotide probes needed for a meaningful analysis and the restriction to serial hybridizations. For probe to target hybridizations the optimal yield of information is gained if the hybridization probability of a

[2] Max-Planck-Institute for Molecular Genetics plasmid library #609

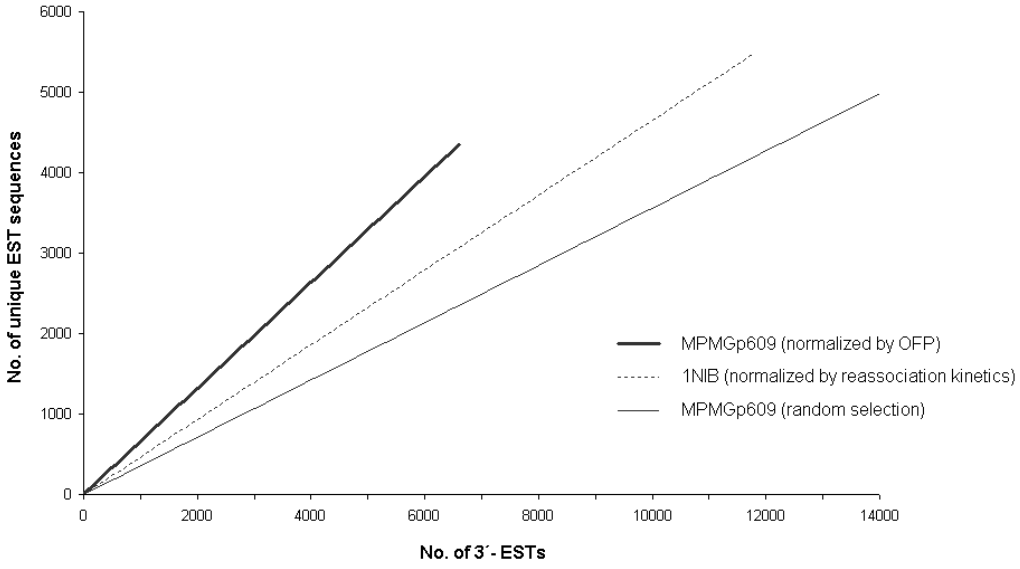

Figure 3.3 Impact of normalization on the redundancy of EST projects

The number of unique EST sequences as a function of the overall number of 3´-ESTs is shown. MPMGp609 (normalized by OFP): more than 6000 3´-sequences of the normalized zebrafish library (Clark et al., 2001) were assembled. Starting from 200 sequence reads, all unique EST sequences were counted by following the chronological order during sequencing. 1NIB[3]: the assembly of roughly 12,000 3´-ESTs of the infant brain library normalized by reassociation kinetics (Bonaldo et al., 1996, Hillier et al., 1996) was carried out in the same way as for the normalized MPMGp609 library. MPMGp609 (random selection): on the basis of the clustering analysis for the whole oligonucleotide fingerprinted MPMGp609 library (~ 21,000 clones) a random selection was simulated. Out of all selected clones only those representing singletons were considered thus allowing to simulate a non-normalized library. The diagram clearly shows the superior normalization performance of OFP.

probe is 50%. One probe yields maximal information when it partitions a given clone set of *N* random clones into two subsets of equal size, i.e. it hybridizes to one half of the clones but not to the other. Two probes should ideally partition *N* random clones into four and *n* probes into 2^n subsets of equal size. In OFP experiments, normally 50,000-100,000 clones are characterized. Assuming a hybridization probability of 50%, about 17 probes would be needed for a successful partitioning of 100,000 clones. In practice, however, hybridization frequencies of the currently used DNA octamer probes are much lower[4]. Consequently, the overall number of hybridized oligonucleotides has to be increased (up to 250) in order to achieve a sufficient partitioning and thus clustering quality of the library to be analyzed (Herwig et al., 1999 and 2000). In view of this high probe number, the

[3] normalized infant brain library #1

[4] Hybridization frequencies of randomly chosen octamer probes to random double-stranded target sequences, that are independent from each other and of equal length of 1000 bp, are ~ 3% (Herwig et al., 2000).

restriction to sequential probe hybridizations leads to a laborious multi-step hybridization procedure that is time- and resource-consuming. In addition to that, the use of radioactive labeling renders the practical implementation hazardous and requires appropriate laboratory facilities for radioisotope handling. Although a novel OFP protocol was recently proposed (Guerasimova et al., 2001) that avoids radioactive labeling by making use of fluorescence as a means of hybridization monitoring, the bottleneck of serial hybridizations still persists.

3.4 MALDI-TOF mass spectrometry

Mass spectrometry is an extremely powerful method for the characterization of compounds. By measuring mass-to-charge ratios of gas-phase ions it is possible to precisely determine the molecular mass of a molecule under investigation and hence to unambiguously identify it. In particular, matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) has revolutionized the mass spectrometric analysis of biomolecules (Karas and Hillenkamp, 1988). Karas and Hillenkamp discovered that irradiation of crystals formed by suitable small and acidic organic molecules, termed “matrix”, with a pulsed laser at wavelengths close to a resonant absorption band of these matrix molecules induced an energy transfer and desorption process, evaporating matrix ions into the gas-phase. More importantly, they found that by incorporation of low concentrated non-adsorbing analytes, such as proteins or nucleic acids, into the crystalline structure of matrix molecules, these analytes were co-desorbed into the gas-phase and ionized upon laser irradiation. The generated ions are then accelerated by an electric field (fig. 3.4). However, the process of ionization in MALDI is not yet well understood. Several mechanisms for ionization of large molecules were suggested (Schlag et al., 1992). Predominantly either positively or negatively single charged molecules are detected (Karas et al., 2000), with these ions being created in a gas-phase proton-transfer reaction with matrix molecules. Usually, MALDI-MS is performed with time-of-flight separation (MALDI-TOF MS) (Hillenkamp et al., 1991). By ion optics molecules are guided into a flight-tube, a field-free drift region, in which they are separated according to their molecular mass before they are finally detected. Improved instrumental designs using delayed extraction have led to mass accuracies of less than 0,01% and significantly improved resolution of signals obtained (Colby et al., 1994, Brown et al., 1995).

MALDI-TOF MS has been applied in different variations for the analysis of proteins, peptides and nucleic acids (Yates, 1998). Analyzing the latter species with MALDI-TOF MS has several advantages. First, its speed of signal acquisition (ionization, size

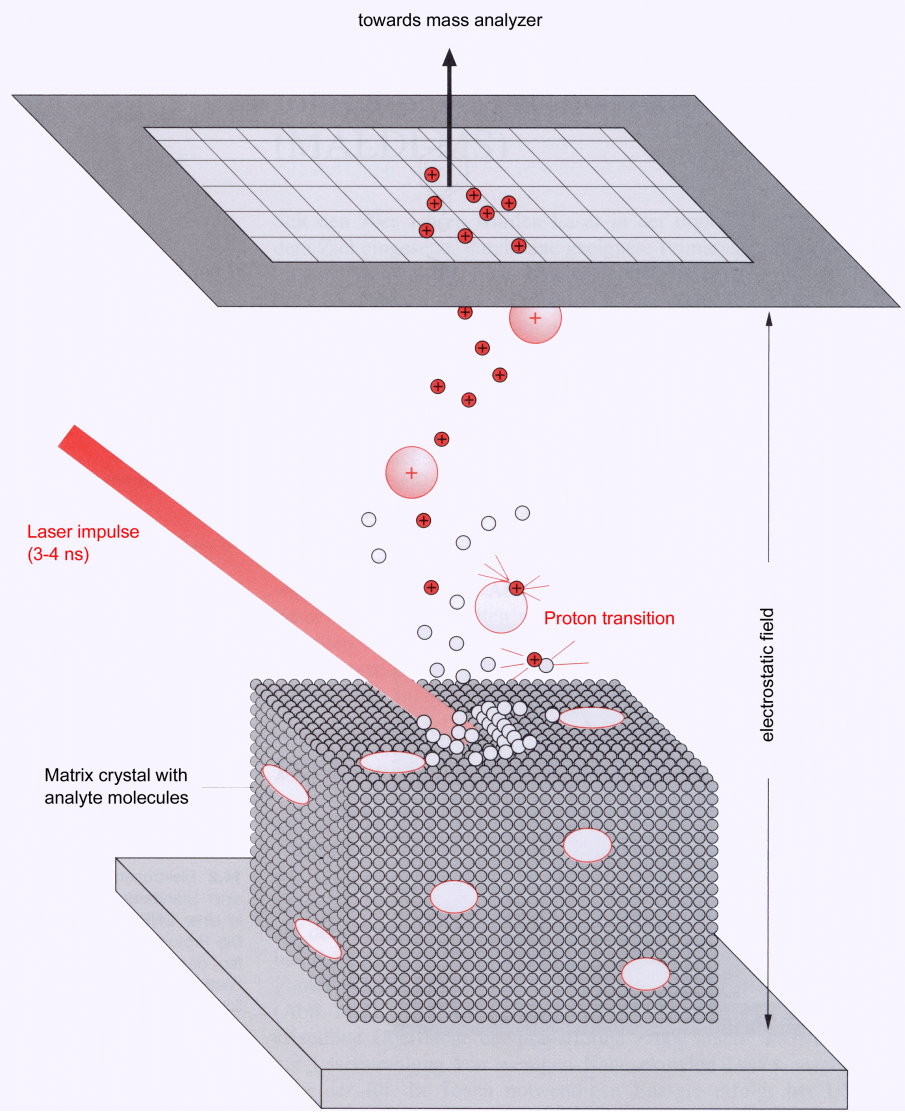

Figure 3.4 Principle of MALDI-TOF MS

Schematic illustration[5] of the process of matrix-assisted laser desorption/ionization. Further details are described in this section.

separation and detection) is very high and lies in the range of milliseconds. Second, the results obtained are highly accurate, because they are based on the intrinsic physical property of mass-to-charge ratio (m/z). In contrast to that, conventional electrophoretic and hybridization array-based methods are both susceptible to complications from secondary structure formation arising in nucleic acids. Furthermore, the absolute nature of detection, i.e. the capacity of analyzing many molecules simultaneously owing to their mass, and the fact that single charged molecular ions predominate, render MALDI-TOF MS compatible for the analysis of complex mixtures of analytes. As a result, it is superior

[5] taken from Lottspeich and Zorbas (1998)

to radioactive and fluorescent means of detection where only one or very few analytes can be detected at a time. Third, the complete automation of all steps, from sample preparation to the acquisition and processing of data, is feasible (Van Ausdall and Marshall, 1998), conferring MALDI-TOF MS great potential for high-throughput nucleic acid analysis applications.

3.5 Peptide nucleic acids as hybridization probes

Oligonucleotides play a central role in a wide range of molecular biology techniques, including DNA/RNA sequencing, DNA cloning, PCR and use as hybridization probes.

The significance of DNA analogues as a tool for hybridization has steadily been grown over the last decade due to the insufficient stability and mismatch discrimination of conventional DNA oligonucleotides. Among the many new compounds introduced as DNA mimicries in recent years a new class of non-ionic oligomers has been of particular interest: peptide nucleic acid (PNA) (Nielsen et al., 1991). In PNA the sugar-phosphate backbone is replaced by an uncharged achiral pseudopeptide backbone, consisting of N-(2-aminoethyl)-glycine units, onto which the nucleobases are attached via a methylene-carbonyl linkage. Owing to its peptide-like structure PNA is chemically much more closely related to proteins or peptides than to nucleic acids (fig. 3.5).

PNA forms very stable duplex hybrids with Watson-Crick complementary DNA, RNA and even PNA oligomers (Egholm et al., 1993. Wittung et al., 1994). Independent studies confirmed that PNA/DNA duplexes have a greater stability compared to DNA/DNA and DNA/RNA duplex hybrids which is ascribed to the electro-neutrality of PNA (Tomac et al., 1996, Jensen et al., 1997). In comparison to their DNA and RNA counterparts, corresponding PNA duplexes show an increased melting temperature (T_m) of 1.0-1.5° C per base. Another advantage is that mismatches generally result in greater destabilization of the duplex reflected by a drop in T_m. However, more systematic investigations of the chemical and positional effects of mismatches on DNA/PNA duplex stability revealed the remarkable stability of some hybrids over others (Weiler et al., 1997). By in-gel affinity electrophoresis, Igloi (1998) demonstrated that for an PNA 11-mer the destabilization effect was at maximum if a mismatch occurred at position four from either end, but negligible in the case of the central position. Furthermore, it was discovered that the most stable mismatched pairings were G:T and T:T whereas maximal destabilization was found for A:A and G:G.

In contrast to DNA, achiral PNA hybridize both in parallel and antiparallel orientation, with the N- to C-terminal fashion (corresponding to the classical DNA 5´-end to 3´-end

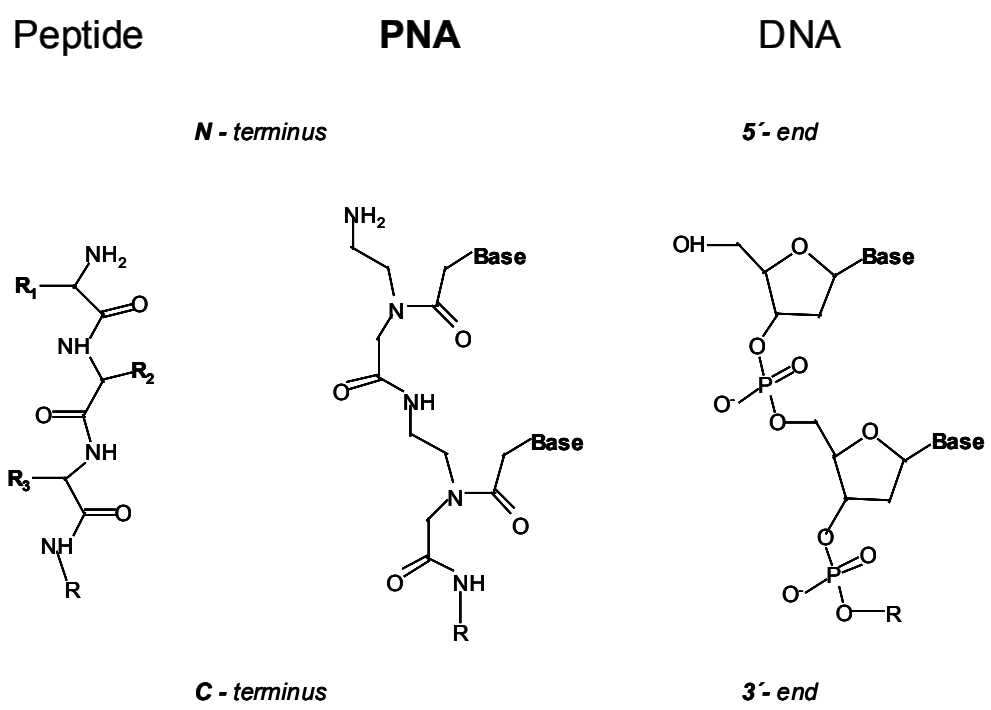

Figure 3.5 Chemical structure of PNA

The structure of a PNA oligomer compared to the structure of a peptide and naturally occurring DNA. The N-terminus of PNA corresponds to the 5´-end of DNA whereas the C-terminus represents the counterpart of DNA´s 3´-end.

orientation) being favored (Egholm et al., 1993, Pfeffer et al., 1993). Due to the lack of electrostatic repulsion, duplex formation is essentially ionic strength independent which results in the ability of PNA to perform strand invasion (Cherny et al., 1993) and enables hybridization under low salt conditions (Orum et al., 1995). At these conditions target DNA is less prone to formation of secondary structures (e.g. hairpins) and thus more accessible to hybridization probes.

These above described properties of PNA have been exploited in a wide range of applications, such as PCR clamping (Orum et al., 1993), affinity capturing (Orum et al., 1995), *in situ* hybridization (Lansdorp et al., 1996), pre-gel hybridization techniques (Perry-O`Keefe et al., 1996), PNA-based biosensors (Wang et al., 1996), and inhibition of gene expression (Good and Nielsen, 1998). A very powerful means of analysis for PNA and PNA hybridization is MALDI-TOF MS, which has been successfully employed for the verification of PNA purity and success of synthesis (Butler et al., 1996) as well as for genotyping of single-nucleotide polymorphisms (SNPs) (Griffin et al., 1997, Ross et al., 1997). PNA in turn is highly suitable for MALDI-TOF MS applications due to its extraordinary stability under most MALDI conditions. In contrast to DNA, PNA due to its uncharged backbone does not fragment easily during the process of MALDI. Furthermore,

it is readily ionized with almost no multiply charged species occurring and does not tend to form metal ion adducts (ion plus sodium or potassium) reducing peak resolution (Butler et al., 1996). The feature of strong binding to complementary DNA under minimal ionic strength conditions renders PNA even more attractive and compatible for MALDI-TOF MS applications.

3.6 A brief introduction to DNA microarrays

Microarrays are one of the most popular areas in today´s biological research. In general, they are monolithic, flat surfaces that bear hundreds to thousands multiple probe sites. Each of these is loaded with a reagent whose molecular recognition of a complementary molecule can lead to a signal that is detected by an imaging technology, most often fluorescence.

The term “DNA microarray” or “DNA chip” refers to the systematic arrangement of DNA molecules on such a surface (e.g. glass, metal, or silicon wafer). DNA microarrays are of topical interest owing to their promising ability for obtaining information on sequences faster, simpler and cheaper than traditional methods. Not surprisingly, major applications of DNA microarrays are projected in large-scale gene expression profiling. Indeed, the parallel processing power of DNA microarrays has encouraged a novel approach to science, stemming from genome sequencing: systematic large-scale investigation, in contrast to hypothesis-driven experimentation.

DNA microarrays exist in two different variants. In the first format, an array of oligonucleotides is synthesized either by conventional synthesis followed by on-chip immobilization or *in situ* by adapting semiconductor photolithography – also known as the "Affymetrix method" of microarray production (Fodor et al., 1991, Pease et al., 1994). The array is subsequently exposed to labeled sample DNA, hybridized and complementary sequences are determined. In the second format, biologically-derived DNA, predominantly in the form of PCR products, is immobilized on a solid surface and exposed to a set of labeled probes either separately or in a mixture. If needed, appropriate functionalization (specific reactive chemical groups or labels) is introduced during the process of PCR. This latter kind of microarray production is the most widely used nowadays. It was crucially pioneered by two groups at Stanford University, USA, (Schena et al., 1995, Lashkari et al., 1997) and has created a big demand for supply that is now commercially met.

Much effort has been dedicated to the sourcing and functionalization of the surfaces on which a microarray is made. A vast amount of different surfaces and attachment chemistries has been applied of which, owing to accessibility, the standard glass

microscope slide has been most widely used. However, more sophisticated surfaces, including fused silica, gold and silicon wafers, are also available. Such DNA immobilization systems comprise epoxy-functionalized glass slides (Eggers et al., 1994, Beattie et al., 1995) or SiO_2-coated silicon wafers (Lamture et al., 1994), self-assembled monolayers (SAM) of thiols (Jordan et al., 1997, Thiel et al., 1997, Brockman et al., 1999), disulfide surfaces (Smith et al., 2001), and sulfanylsilane-derivatized glass (Rogers et al., 1999) as well as novel modifications of silicon (Strother et al., 2000a und 2000b) or positively charged polylysine surfaces (Sinibaldi et al., 2001). Three-dimensional innovations have been the use of dendrimeric linkers (Beier and Hoheisel, 1999, Benters et al., 2001) and photolitographically produced gel-pads (Yershov et al., 1996). More exotic applications were reported by Livache et al. (1994 and 1998), Gilles et al. (1999), and Heller et al. (2000) making use of electropolymerization and electric field, respectively.

With many methods available for DNA microarray production, the strengths and weaknesses of each have to be individually considered. Arrays spotted from cDNA and those derived from synthetic oligonucleotides (either synthesized and spotted or synthesized *in situ*) have significantly different applications. The remaining question whether there is a preferred method for microarray production cannot be answered without knowing the individual needs and resources.