Molecular Taxonomy. Bioinformatics and Practical Evaluation

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Alexander Pozhitkov

aus Moskau, Russland

(Köln, 2003)

Berichterstatter:

Prof. Dr. Diethard Tautz

Prof. Dr. Thomas Wiehe

Tag der mündlichen Prüfung: 02. December 2003

Acknowledgements

I very much grateful to my supervisor Prof. D. Tautz for the opportunity to join his group and satisfy my passion to the research. I am also grateful to him for giving me freedom and at the same time a delicate guidance throughout my work. I would like to thank Prof. T. Wiehe, Prof. D. Schomburg and Dr. R. Wünschiers for accepting the membership in my theses committee.

My best friend Tomislav Domazet calmed me down many times and helped me to be realistic and sober concerning my results and approaches. Our long discussions brought a lot of fruits into my work. I am thankful to Hilary Dove, her kindness and support.

I am grateful to Dr. Lysov from the Engelgardt Institute of Molecular Biology, Russian Academy of Sciences for the supporting in the initial phase of the project. I thank Prof. Speckenmeyer at the Institute of Informatics, University of Cologne for providing access to their LINUX cluster and J. Rühmkorf for his help with installing the parallel version. D. Ashton (Argonne National Lab) greatly helped me with the windows version of the MPI. I would like to thank to Dr. M. Gajewski for his help with establishing of the microarrays.

I would like to specifically show gratitude to Dr. H. Fusswinkel for her help with some very complex administrative issues. Greatly appreciated help from E. Sigmund and G. Meyer.

I am particularly thankful for my mother and my wife for the encouragement. My father greatly helped me scientifically to clarify many technical questions of my work.

This work was supported by a grant from the Ministerium für Schule Wissenschaft und Forschung des Landes Nordrhein-Westfalen.

Abbreviations

DNA	deoxyribonucleic acid
RNA	ribonucleic acid
rRNA	ribosomal ribonucleic acid
CPU	central processing unit
GUI	graphic user interface
OS	operation system
PC	personal computer
DIY	do-it-yourself

Zusammenfassung

Mit Hilfe der molekularen Taxonomie wird die biologische Diversität von Organismen anhand von molekularen Markern untersucht. In dieser Arbeit wird eine Methode entwickelt, um kleine Organismen durch molekulare Taxonomie zu charakterisieren. Da die Nukleotidsequenzen Ribosomaler RNA (rRNA) Regionen aufweist, die verschiedene Ebenen der Konservierung haben, können sie als Art-, Genus- oder Taxonspezifische molekulare Marker dienen.

Die Organismen leben in komplexen Ökosystemen. Um die Artenzusammensetzung dieser Ökosysteme zu untersuchen, wurde ein Hybridisierungsansatz mit Oligonucleotid Microarrays entwickelt um das Vorhandensein einer bestimmten rRNA aufzuzeigen. Zusätzlich wird hier ein zweiter Ansatz auf der Basis der Pyrosequenzierungtechnologie vorgestellt. In diesem Fall wird eine Mischung von rRNA Molekülen direkt sequenziert und der Anteil der einzelnen Arten wird dann von dem erhaltenen Pyrogram errechnet.

Diese Arbeit lässt sich in zwei Teile geliedern: theoretische Bioinformatik und experimentelle Ansätze. Der erste Teil befasst sich damit, die Stabilität der DNA/RNA Duplexe vorherzusagen. Als Ergebnis wird eine *ad hoc* Stabilitätsformel vorgestellt. Ein Algorithmus und ein Program wurden entwickelt, um Oligonucleotide für den microarray Ansatz zu entwerfen. Ausserdem wurden die kinetischen Aspekte der Dissasoziation der DNA/RNA Duplexe berücksichtigt. Zusätzlich wurde der Formalismus des Pyrosequenzierungs Ansatzes theoretisch bearbeitet.

Die experimentelle Teil befasst sich mit den Einzelheiten der Oligonucleotid Microarray Technologie, unter anderem mit der Herstellung der Arrays, Immobilisierung, Hybridisierung und mit dem Scannen. Ein "real-time" kinetischer Aufbau für die Beobachtung der DNA/RNA Duplex Dissasoziationen wurde entwickelt. Die theoretischen Ergebnisse und die Qualität des Oligonucleotiddesigns wurden praktisch ausgewertet, und es wurde festgestellt, dass die Theorie den experimentellen Ergebissen gut entsprach. Der Pyrosequenzierungsansatz wurde auch getestet und es wurde gezeigt, dass angewandt werden kann um die Zusammensetzung einer komplexen Mischung von rRNA Genen festzustellen.

Summary

Molecular taxonomy is a field that studies the diversity of organisms based on molecular markers. This work is devoted to develop a methodology of molecular taxonomy of small organisms. The ribosomal RNA (rRNA) is used as a molecular marker since its nucleotide sequence includes stretches of various levels of conservation, which can be used as species, genus and taxa specific regions.

The organisms live in complex communities. To discover the composition of these communities, a hybridization assay employing oligonucleotide microarrays is developed to indicate the presence of a certain rRNA, in a sample under investigation. An additional method based on the pyrosequencing process is proposed here. In this case the mixture of rRNA genes is directly sequenced and the proportion of individual sequences is then calculated from the obtained pyrogram.

The work comprises two parts: theoretical bioinformatics and practical evaluation. The first part tackles the problem of DNA-RNA duplex stability prediction. As a result, an *ad hoc* stability function is proposed. An algorithm and a program are developed for the design of oligonucleotides employed in the microarray approach. The kinetics of DNA-RNA duplex dissociation is considered as well. In addition, the formalism of the pyrosequencing approach is elaborated theoretically.

The experimental part deals with the issues of oligonucleotide microarray establishment, including fabrication, immobilization, hybridization and scanning. A real-time kinetic setup for observing the RNA-DNA duplex dissociation was developed. The theoretical findings and quality of the oligonucleotide design are practically evaluated. The theory is found to be in a good accordance with experiment. The pyrosequencing approach is tested as well and is demonstrated to have enough power to discover the composition of a complex mixture of rRNA genes.

Introduction

Molecular taxonomy is an appealing way of studying the ecology of small organisms without cultivation and visual determination. A key to the molecular taxonomy is the fact that each organism contains ribosomes, and that their structural RNAs on the one hand have enough diversity to be unique for a particular species, on the other hand possess conserved regions common for all taxa. The identification of species or species groups with specific oligonucleotides as molecular signatures is becoming increasingly popular for bacterial samples. However, it shows also a great promise for other small organisms that are taxonomically difficult to tract. DNA microarrays are currently used for gene expression profiling [1, 2], DNA sequencing [3], disease screening [4], diagnostics [5, 6], and genotyping [7], usually within the context of clinical applications. The extension of microarray technology to the detection and analysis of 16S rRNAs in mixed microbial communities likewise holds tremendous potential for microbial community analysis, pathogen detection, and process monitoring in both basic and applied environmental sciences [8-10]. There are several types of microarrays available on the market and the oligonucleotide microarrays are among them. The work here solely deals with oligonucleotide microarrays, both theoretically and practically. Two major problems that have been addressed in this work are: (i) design of the optimal oligonucleotide with desired specificity and (ii) practical evaluation of designed probes.

I have devised here an algorithm that aims to find the optimal probes for any given set of sequences. The program requires only a crude alignment of these sequences as input and is optimized for performance to deal also with very large datasets. The algorithm is designed such that the position of mismatches in the probes influences the selection and makes provision of single nucleotide outloops. Program implementations are available for Linux (text version) and Windows (text and GUI version). The soundness of the results produced by the program has been tested experimentally.

In addition, a microarray free approach based on sequencing of a mixture of genes has been developed in this work. The microarray free approach makes use of a novel pyrosequencing method and discovers the mixture composition quantitatively. Here

only the principle is proven and the approach has been tested on the artificial mixture of DNA encoding rRNA.

The work contains five chapters. The first chapter deals with the bioinformatics of a probe design. The second chapter depicts a new paradigm of the graphic user interface strategy applied to the probe design. The third chapter is mainly devoted to the technical establishment of the microarrays. The fourth chapter experimentally evaluates the probe design. Finally the fifth chapter deals with the development of the microarray free method.

Chapter 1 An Algorithm and Program for finding Sequence Specific Oligo-nucleotide Probes for Species Identification

Introduction

Identification of species with molecular probes is likely to revolutionize taxonomy, at least for taxa with morphological characters that are difficult to determine otherwise. Among these are the single cell eucaryotes, such as Ciliates and Flagellates, but also many other kinds of small organisms, such as Nematodes, Rotifers, Crustaceans, mites, Annelids or Insect larvae. These organisms constitute the meiofauna in water and soil, which is of profound importance in the ecological network. Efficient ways for monitoring species identity and abundance in the meiofauna should significantly help to understand ecological processes.

Molecular taxonomy with sequence specific oligo-nucleotide probes has been pioneered for bacteria [10,11]. Probes that are specific to particular species or groups of related species can be used in fluorescent in situ hybridization assays to detect the species in complex mixtures or as symbionts of other organisms [12,13]. Alternatively, the microarray technology is increasingly used for this purpose, allowing potentially the parallel screening of many different species. Most of the species-specific sequences that are used so far for this purpose are derived from ribosomal RNA sequences. However, any other sequence is also potentially suitable, as for example mitochondrial D-loop sequences in eucaryotes.

The species-specific probes are usually derived from an alignment of the respective sequences, where conserved and non-conserved regions are directly visible. A program has been developed for ribosomal sequences that helps to build the relevant database, and supports the selection of suitable specific sequences (ARB [14]). In this, a correct alignment is crucial for finding the optimal probes, but alignments are problematical in poorly conserved regions. These, on the other hand, have the highest potential to yield specific probes. Moreover, the current implementation of probe finding calculates only the number of mismatching position to discriminate between the probes, but does not take into account the position of the mismatches within the stretches, which could influence the hybridization behavior.

We have therefore devised here a new algorithm that allows working with datasets that need not to be carefully aligned and that takes the position of mismatches along the recognition sequence into account.

The algorithm

The algorithm includes three parts. The first one aims to provide a function that calculates the relative stability of matching oligos in dependence of the number and position of mismatches. The second one provides a strategy for probe finding that scans all possible sequence combinations, but works time efficient. The third part deals with matches caused by single nucleotide outloops of a given sequence.

Stability function

Extensive studies exist for assessing the thermodynamic consequences of internal mismatches in short oligo-nucleotides (see fro example [15,16]). These show that there are no simple rules and that the exact influence on the stability of a hybrid depends on the nature of the mismatch, as well as its flanking nucleotides. For example, mismatches including a G (i.e. G-G, G-T and G-A) tend to be less destabilizing than the other types of mismatches [16], although this can not directly be predicted from steric considerations. Comparable systematic studies on the relative influence of the position of the mismatch within the oligonucleotide do not exist yet, although it is clear that the influence is lower at the ends than in more central positions [16, 18]. Preliminary evidence with an oligo-dT stretch harboring A mismatches along the sequence suggests that the position dependence could be a continous function [17]. We have therefore decided to use an *ad hoc* approach for the stability calculation that is mainly designed to discriminate against sequences with more central mismatch positions.

We model the relative stability of mismatched oligos as follows. The position of the mismatch can be considered to be a "weak point". The location of the "weak point" is expressed as a probability function that takes into account the differential contribution of central versus terminal positions. The probability that the "weak point" is at position x is defined by p_1. Under the experimental conditions of melting, the presence of the "weak point" is true, meaning that [sum(p1) for all x] = 1.

We assume a Gauss distribution as the respective probability function, with the maximum in the middle of the duplex and the integral value along the duplex length set to 1 (Equation 1-1).

$$p_1(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\frac{L-1}{2})^2}{2\sigma^2}};$$

$$\sum_{x=0}^{L-1} p_1(x) = 1.$$

Equation 1-1. "Weak point" location probability. L – duplex length, σ - distribution parameter, x – duplex position.

Note that the function in Equation 1-1 refers to discrete positions within the sequence, while the Gauss distribution is continuous and the integration from –∞ to +∞ is set to yield 1. The parameter σ is therefore chosen such that the discrete sum approaches 1 at any intended precision. In the program discussed below the accuracy of the sum value is 0.999.

Although the preliminary experimental evidence [17] suggests that the destabilization function can be approximated with the Gauss distribution, the program implementation allows also to use a flat distribution, i.e. where a position-independent effect on the melting is assumed as an alternative, to compare the outputs of the two different assumptions.

For assessing the relative amount of destabilization caused by a certain mismatch, we assume that the mismatch disturbs the surrounding basepairs from (y-n) to (y+n) positions. n can be called a border parameter that will need to be experimentally verified in the future. Because n can currently only be guessed, it is set as a program variable with a default value of 5. n might also depend on the nature of the mismatch, i.e. some types of mismatches might influence the surrounding bases less than the others. We therefore implemented further program variables that allow to define a different n depending on the nature of the mismatch (i.e. it is possible to set a particular n value for each possible type of mismatch).

The overall relative stability of a given duplex is then expressed as a probability function. It is expressed as the sum of products of the individual position probabilities p_1 (determined by the stability function) and p_2 (determined by the border parameter).

The value of p_2 it the probability of "melting", conditioned that the “weak point” is disturbed. (Equation 1-2).

$$\sum_{0}^{L-1} p_1 p_2 = p$$

Equation 1-2. L – the length of the duplex, p1 – the “weak point” location probability, p2 – the "melting" probability due to the disturbance of the “weak point”.

p_2 is a conditional probability of "melting" with $p_2 = 1$ if the “weak point” is disturbed (in the region $y \pm n$) and $p_2 = 0$ at non-affected positions. This allows transforming Equation 1-2 into Equation 1-3.

$$\sum_{y-n}^{y+n} p_1 = p$$

Equation 1-3. y – the mismatch position, n – the border parameter

p_1 can then be substituted by the function in Equation 1-1, to yield Equation 1-4.

$$\sum_{y-n}^{y+n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\frac{L-1}{2})^2}{2\sigma^2}} = p$$

Equation 1-4. x – the duplex position, y – the mismatch position, n - the border parameter

In the case of several mismatches, the summing is done along all the respective mismatch regions. If the mismatches occur next to each other, their disturbed regions simply overlap and the summing is performed across the respective region.

Probe finding

The probe finding strategy is devised in a way (i) to avoid the need for exact alignments, (ii) to check probe specificity along the whole available sequence and (iii) to optimize performance. The workflow is depicted in Figure 1-1.

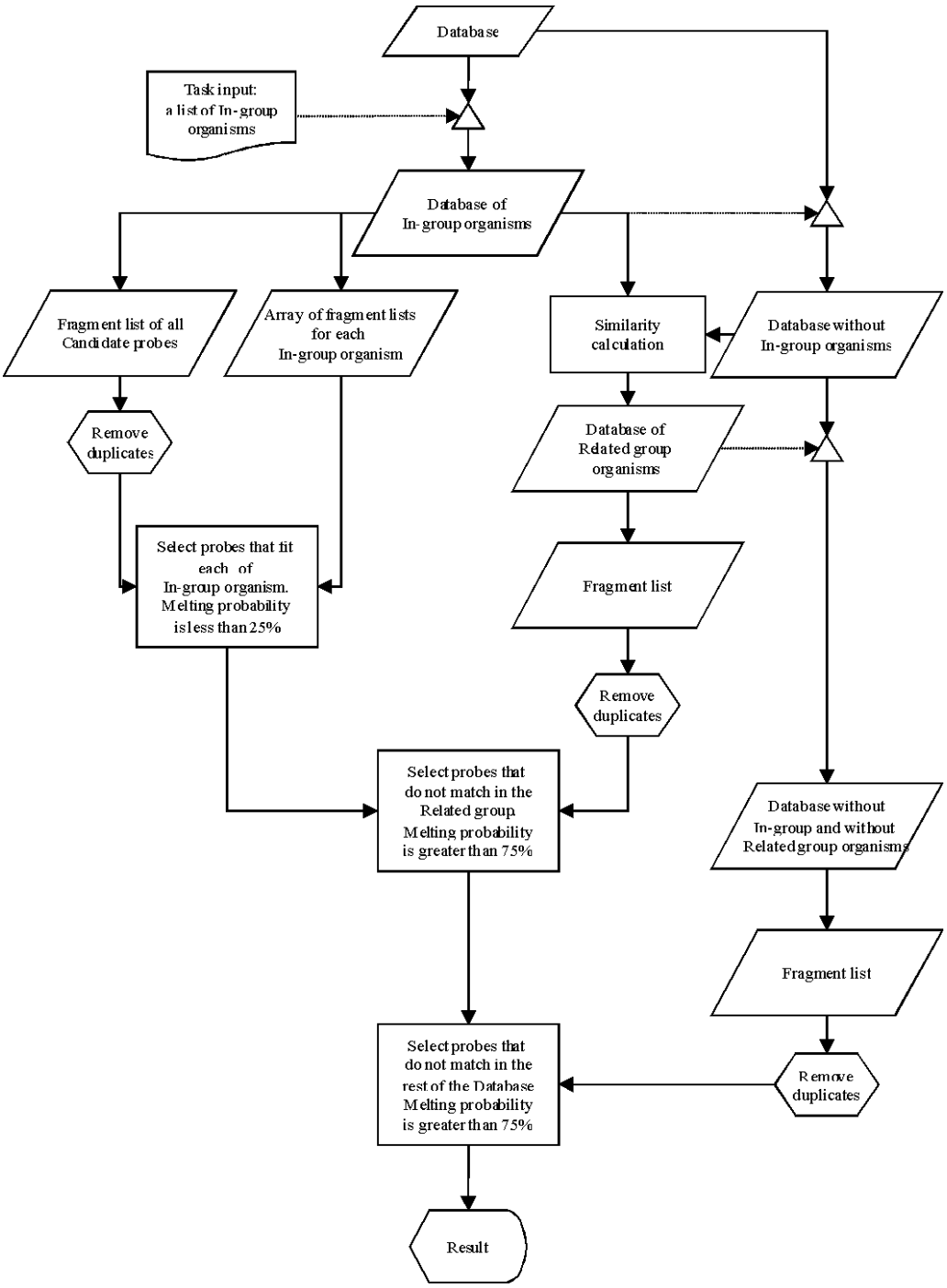


Figure 1-1. Scheme of the probe finding algorithm. Details are explained in the text.

It starts with a database in which each organism is represented by a single continuous sequence, such as a defined region of the 18S or 28S ribosomal genes. From this it takes first the sequences of the In-group organism(s) for which specific probes should be found and cuts these into short pieces of the specified oligo-nucleotide length (set as a program variable), following an approach proposed by Bavykin et al [20]. This is accomplished by a sliding window scheme with 1-nucleotide shifts across the whole length of the sequence(s). Two separate lists are

created in this way. The first list is simply a straight list of all possible fragments from all In-group organisms. The second one consists of an array of lists for each of the In-group organisms (the two lists are identical if only one In-group organism is chosen). All duplicate oligos from the first list are then removed and each of the remaining oligos is checked whether it matches with each of the In-group organisms in the second list. A match is positive, when the relative melting probability is within the range of 0 - 25%, employing the function of Equation 1-4. Thus, this first calculation simply ensures that all candidate probes match with all In-group organisms. This calculation would be largely dispensable, if only a single In-group organism is chosen.

The next step is to subtract all oligos that match in any of the Out-group organisms. To avoid the comparison of all candidate oligos against all Out-group sequences, we identify first a group of sequences that is closely related to the In-group. For this one requires a rough alignment of all sequences, to calculate percentage similarity between them. Note that this serves only to identify a subgroup of sequences for speeding up the calculations, i.e. mistakes in the alignment are of no concern. The similarity calculator in the program extracts this related group of sequences by a simple percentage identity calculation across the given alignment. All sequences that are at least 90% similar to the In-group are used as Related-group. This percentage can be set as a program variable and should be set such that the Related-group does not become more than 5-10% of all sequences.

The sequences of the Related-group are again converted into a fragment list as above, duplicates are removed and all candidate oligos are matched with this list. Now only those oligos are retained, which have a melting probability of at least 75% (the exact percentage values are program variables). The majority of oligos is removed in this step. The remaining candidate oligos are then matched against the remaining sequences in the Out-group with the same cut-off criterion.

This stepwise selection scheme allows to significantly speed up the calculations even for very large datasets, but still ensures that all oligo-nucleotides of the desired length were directly or indirectly matched against all possible other oligos in the database.

Single nucleotide loops

Structure analysis with experimental oligo-nucleotides has shown that in a pair of hybridized oligos, one nucleotide can loop out, without interfering much with the stability of the hybridized pair [21]. This implies that one base of one strand of a duplex can loop out from the duplex and the rest of the strand can shift one position. This is depicted in Figure 1-2.

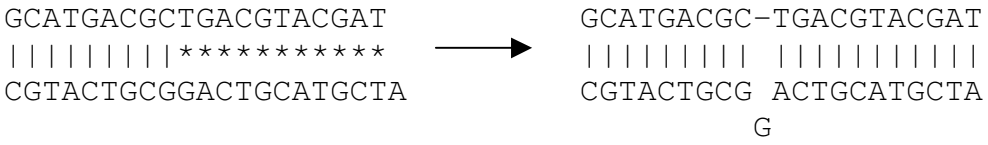


Figure 1-2 Scheme of the single-nucleotide outloop problem; asterisks represent mismatches, columns represent matches.

A standard linear scanning algorithm would recognize the situation at the left as one with 11 mismatches, i.e. would suggest it as a specific probe. However, if the single nucleotide loop is taken into account, the match would be perfect and the probe would have to be considered as unspecific. Our scanning algorithm takes this problem into account by re-checking all candidate probes after the completion of the filtering steps. It does this by sequentially removing one nucleotide from the candidate probe and shifting the remainder by one position. The melting probability of the new oligo is then calculated and checked. The removed nucleotide is then reinserted and the cycle is repeated for the next position. The same procedure is done for the target sequence, so that outloops are considered to be possible on both strands of the duplex. Note that outloops of two nucleotides are considered to destabilize the helix too much to warrant a separate analogous calculation.

Parallel computation

A parallel program version allows probe finding to be done in parallel on several processors. Essentially the same algorithm is used in the parallel version of the program, whereby the parallelism is introduced in the matching steps. Each process takes its own part of the database and performs the matching as well as the stability calculations. The results are then gathered by the root process and superimposed.

Program implementation

The algorithm is implemented in a program called PROBE. The program consists of three modules that can be used independently. The first module finds the probes based on the given task (specificity group, length of probes, source database).

The second one is the analytic module, which can be used if it is impossible to design a probe for a given organism group. This module depicts the situation with the given In-group and enables to find the closest group for which the task can be accomplished. The use of the analytic mode comes into play when PROBE fails to identify a set of probes for the given organism group. Such a failure can have two reasons - either there is no probe, which identifies all organisms in the specificity group, or there is another organism outside the specificity group, which is also identified by all candidate probes suitable for the specificity group.

For the first case, the specificity group must be broken down into several subgroups and the probes must be identified for these subgroups separately. For the second case, the organism that is very similar to the specificity group should be added to the specificity group and this may then have to be broken down into smaller subgroups.

The analytic module creates a table with the organisms of the specificity group as well as the most related organisms. This table depicts then the matching or non-matching patterns for each of the possible probes, allowing a simple visual inspection of the best specificity groups. The output can be viewed and modified with spreadsheet programs such as Excel.

The third module provides a report for the identified probe, including the mismatches in the duplexes within the specificity group, the best match out of the group and some other information.

The program is written in standard C++ in a platform independent manner. Therefore, the program can be easily compiled for Linux and Windows without any modifications. The program binary files for Linux and Windows are available from the http://biochip.genetik.uni-koeln.de/probe as freeware accompanied with all its source files, and a manual that describes further details.

Results

As an example of the performance of the program we have used the full SSU database (RDP, release 8.1) [22] containing approximately 16.000 sequences to find a specific oligo-nucleotide probe with a length on 20 nt for *Thermotoga maritima*. The search was done on a Pentium III (800 MHz, 512 MB RAM) PC and took about 1.5 hours without outlooping and 16 hours with outlooping, indicating that the most time intensive step is the outlooping subroutine. The parallel version running on a cluster with 24 nodes (with the slowest node being a Penthium II - 400 MHz with 256 MB RAM) took 2 hours for the same full task.

Figure 1-3 depicts the output from the check module, which allows comparing the oligos and their specificity that were found in this particular comparison. It shows that ARB suggests two oligos that are rejected by PROBE either because of mismatches occurring only at the ends, or under the outloop routine. Both programs find one oligo with acceptable high specificity.

A

```
Target:
477    AAACCCUGGCUAAUACCCCA
Probe:
tggggtattagccagggttt
Ingroup, matching:
Duplex:
477    AAACCCUGGCUAAUACCCCA  Thermotoga maritima str. MSB8 DSM 3109 (T).
477    AAACCCUGGCUAAUACCCCA
melting probability   0

Outgroup, matching (without outloop):
Duplex:
1200   UGGCCCUGGCUAAUACCCGGG Ralstonia eutropha str. DS185.
477    aaaCCCUGGCUAAUACCCca
melting probability   0.42

Outgroup, matching (outloop)
Duplex:
477    AAACCCGGCUAAUACCGCAUA Thiorhodovibrio sp.
477    AAACCCGGCUAAUACCcCA outloop: 6
melting probability   0.30
```

B

```
Target:
1143   AAACCGCUGUGGCGGGGGAA
Probe:
ttcccccgccacagcggttt
Ingroup, matching:
Duplex:
1143   AAACCGCUGUGGCGGGGGAA  Thermotoga maritima str. MSB8 DSM 3109 (T).
1143   AAACCGCUGUGGCGGGGGAA
melting probability   0

Outgroup, matching (without outloop):
Duplex:
571    GCCCUGCUGUGGCGGGGUCAG Treponema uncultured Treponema clone RFS60.
1143   aaaCcGCUGUGGCGGGGgaA
melting probability   0.75

Outgroup, matching (outloop)
Duplex:
570    GGCCCGCUGUGGCGGGGUCA outloop: 5      Treponema clone RFS60.
1143   aaaCCGCUGUGGCGGGGgaA
melting probability   0.509097
```

C

```
Target:
1265   ACGGUACCCCGCUAGAAAGC
Probe:
gctttctagcggggtaccgt
Ingroup, matching:
Duplex:
1265   ACGGUACCCCGCUAGAAAGC  Thermotoga maritima str. MSB8 DSM 3109 (T).
1265   ACGGUACCCCGCUAGAAAGC
melting probability   0

Outgroup, matching (without outloop):
Duplex:
1731   GAAGCGCCCCGCUAGAACGCG Sulfolobus solfataricus str. P1 DSM 1616 (T).
1265   acgGuaCCCCGCUAGAAaGC
melting probability   0.88

Outgroup, matching (outloop)
Duplex:
1264   GAGCGUACCCGCUAGAAAGC outloop: 10     clone WCHB1-64.
1265   acGguacCCCGCUAGAAAGC
melting probability   0.74
```

Figure 1-3 Comparison of specific oligos suggested by ARB and PROBE for *Thermotoga maritima*, in comparison to the whole SSU database. A) **Oligo suggested by ARB, but found to have lower than 70% melting probability in two other species. This was therefore rejected by PROBE because of insufficient specificity.** B) **Oligo suggested by ARB, but found to have lower than 70% melting probability when outlooping is considered. This was therefore also rejected by**

PROBE because of insufficient specificity. C) **Oligo suggested by both programs, whereby the best outgroup matches have a higher than 70% melting probability.**

Discussion

The algorithm presented here does not take into account the effect of relative GC content and stacking interactions of neighboring bases on the melting temperature of the oligo-nucleotides. Accordingly, the oligo-nucleotides suggested by the program can differ significantly in melting temperature. However, as this can easily be adjusted after the selection is made, we have not included a subroutine that takes GC content into account during the primary search, because this would slow down the calculations. Furthermore, we expect that GC content differences may be of less importance for the applications envisioned here, because they can be largely compensated by the choice of experimental conditions, such as buffers that compensate stability differences [22].

A more general problem is our way of calculating the relative stability factor. This does currently not take the nucleotide composition into account either. The reason is that there are too few experimental data as yet, that would allow to unequivocally include this in the calculations. The current experimental data sets focus on the types of mismatches in particular contexts, but not systematically on position specific effects [16, 24]. Moreover, they deal with relatively short model oligos only (up to 12 nt). However, the probes used for species identification are longer and the different effects can currently not be accurately assessed from experimental data for such longer probes. In our equation, it is mainly the border parameter n that would be affected by base composition and nearest neighbor interactions and we have therefore left this as a variable that can be set according to experimental results. In principle, it seems possible that n differs for different sequence compositions, i.e. GC-rich stretches have a smaller n than AT-rich ones. Thus, if one chooses a low n, one would risk that GC-rich oligos are suggested as specific probes that still show cross hybridization. However, it seems that these can easily be eliminated after the selection is made. Still, if experimental data indicate that this is a major problem, the program could easily accommodate such new insights.

Finally, the stability function proposed in Equation 1-1 could possibly also have other shapes than Gaussian. Again this is a factor that needs further experiments. If it

turns out that other functions are more appropriate, one can include this as additional options into the program. At the present we offer the extreme, namely a flat function, as an alternative option.

Conclusion

We have designed a versatile algorithm for finding optimal species- and group-specific probes for molecular taxonomy that is sufficiently open to implement further experimental insights into the nature of the stability of mismatched oligo-nucleotides.

Chapter 2 Graphic User Interface (GUI) for the PROBE. A new Design Paradigm

Introduction

A GUI makes application much more comfortable to work with and more appealing to look at. Unfortunately this is not always true, sometimes even simple applications become complex if the GUI is not well elaborated. This chapter describes a new design and programming paradigm directed towards the user friendly and obvious applications. All discussions and considerations apply Microsoft Windows operating system (OS) and the GUI version of the PROBE is only available for this OS.

The main problem of many graphic applications is their awaiting of the user orders. On the one hand this is of course a desired behavior, if the application is mostly a container of the user's input, for example a text editor or an electronic table. In this case the application is supposed to accept the input and be ready to display it to the user. If the user wants to format the text, perform spell checking and this like, the application has commands for it, and they are usually intuitively clear.

A different situation exists for scientific applications that deal with something essentially new. In this case, the awaiting behavior of the software can be quite confusing, especially if the problems are complex. Many examples of puzzling software are found among academic and commercial products. For instance, upon the startup the application shows a gray window and waits for the user's actions. One of the odd features of many applications is that the menus of hundreds of applications are essentially the same: File, View, Tools, Help. The developers try to split the commands among these four menus, sometimes producing peculiar assignments.

Interestingly, there is no work published that specifically addresses the strategies of GUI design. For example Petzold [25] or Winnick Cluts [26] describe only the facilities for windowing, dialog boxes, progress bars, etc. offered by Windows OS, but does not provide any strategic recommendations on how to use these facilities to create a user friendly GUI. An extensive search within the database of the Institute of Scientific Information (The Thomson Corporation, USA), covering most scientific

journals including the ones devoted to computer science, does not reveal publications specifically dealing with the strategy of designing a GUI.

Owing to the lack of a systematic view on the strategy of GUI design, I propose here a paradigm turning the application from the passive worker towards the active master. Hence, the software but not the user solves the problem. The desired behavior is somewhat similar to that of a "program installation wizard" and various "wizards" that can be found among MS Office applications. In fact, one should understand that changing of the program behavior is not only a question of the design, but a new paradigm. The reason that most of the GUI software is waiting for the user actions lays in the fundamentals of Windows.

Windows Application Fundamentals

Under Windows the applications are "message driven" [25, 27]. Messages are actually the representations of events occurring during the lifetime of an application. For example, the user clicks, pressing on the buttons, pressing on the keyboard, changes of the directory content and many other things are the events. When an event occurs, the internals of Windows generate a message – an integer value specific to each event – and the message is put into the application's message queue. The application is running a loop (so called "message loop") that is getting messages and delivering them to the application. The delivery means invoking a procedure within the code of the application that is a specific response on the particular message. That is why the application is mostly waiting because it is running the message loop and performing any activity only if there is a message to be picked up and processed via the call of its dedicated procedure.

New Paradigm

How to make the GUI application to be not awaiting for the user actions but to be active itself? Apparently there is no way to avoid waiting in the message loop. Fortunately Windows is a multitasking OS that allows to run an application having several threads. A thread is a fragment of code concurrently running with the main application. In fact, the main application is a thread as well. The thread resembles another program coexisting with the main application.

Threads make it possible to separate user interface tasks from the logic of the program. Figure 2-1 shows the multithreaded arrangement of the GUI version of the PROBE. Interestingly, the logic thread of the PROBE is running the same algorithm as was described in Chapter 1.

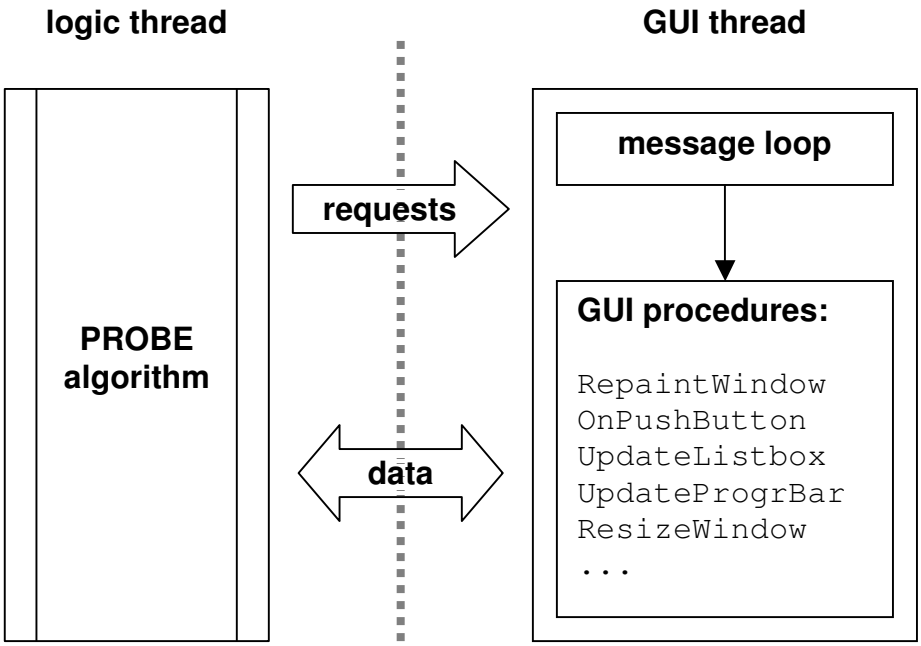


Figure 2-1 Multithreading for GUI. The logic thread is running a text-based algorithm, while the graphic thread is executing typical windowing routines.

In fact, there are no significant changes in the code that has been developed for the text version of the PROBE, only the text input/output statements are changed to the statements making the GUI thread to ask the user either for the input or to display the output. Thus, the paradigm for the creation of the active non-waiting application can be formulated in the following way:

- develop a text based code that is active by default
- create a GUI and organize it as a thread
- create a logic thread and establish inter-thread communication
- insert the text based code into the logic thread and adjust input/output operations.

Implementation

The GUI versions of both single processor and parallel versions of the PROBE have been implemented. According to the new paradigm, the probe calculation engine stays unchanged but a new GUI thread is added. The GUI thread provides means to be queried for input/output, which are in turn the means of inter-thread communication. The probe calculation engine is inserted into the logic thread. The versions have been implemented without the use of Microsoft Foundation Classes (MFC). Although MFC is almost a standard for windows applications, here it was avoided because MFC dictates a very rigid message-driven architecture. Instead, pure Win32 API function calls were used for all GUI tasks.

GUI Objects

The code of the PROBE algorithm is object oriented. The same applies for the GUI part of the application. The object is a programming artifact that in other words is called a "user defined type" or "class" [28]. The C++ classes are language structures that encapsulate its own data and exert methods – procedures that can be called within the program. Each class has a constructor – a method that is automatically called when an object is created in the program code. The GUI objects are in effect C++ classes wrapping the functionality of overlapped windows, dialog boxes and this like. These objects are created by the GUI thread; the constructors contain Win32 API calls that generate windows or dialog boxes and display them. The methods of these objects are of two types. The first type responds on the windows messages delivered by the message loop. These first type methods are not available for invocation from anywhere of the program and dedicated solely for message processing. The examples of such methods are: OnDraw, OnResize, OnDestroy etc. The prefix "On" designates that the method is called upon arrival of the corresponding message. The second type is dedicated for the information exchange with the logic thread.

The logic thread uses the GUI objects through their pointers (addresses in the memory) by invoking the second type methods. These methods are, for example, TextOut, ReadSettings, RetrieveLines. In addition these methods are blocking methods from the perspective of the logic thread. Here the blocking means that the execution of the logic thread is stopped until the data is actually retrieved from the

user, for example, when the method ReadSettings of the DialogBox object is called, it will not return before the user has filled up the fields of the dialog box and pressed "OK" button. Again from the perspective of the logic thread this is pretty much the same as if it was a text-based application, for which in fact the algorithm running in the logic thread has been initially developed. In effect, the logic thread is driving the calculation process and the graphic user interface is entirely dependent on the logic of the algorithm.

Inter-thread Communication

As it has been already mentioned, the threads communicate through the second-type methods of GUI objects. The mechanism of communication employs the user defined Windows messages. Earlier in this chapter it has been stated that the Windows messages are reflections of the events happening during the lifetime of the application, but in fact an application can generate such events by itself and send messages to itself or even to another application. The application can send standard Windows messages or user (in reality programmer) defined ones.

When the logic thread needs some user input or is ready to display the output, it invokes a second type method of the corresponding GUI object. The internals of the method post a standard or a programmer defined message to the message queue of the GUI thread and the method is entering an infinite loop awaiting when the GUI thread completes the request. When the request is completed, it raises a flag that is a signal to quit the loop and return from the method. The pseudo-code below illustrates how the GUI object is arranged and how it participates in both threads simultaneously.

GUI object	
Logic thread	GUI thread
PostMessage(MSG); WaitUntilFlag; return;	OnMSG; RaiseFlagReady;

Technically it is only possible to send a message to a window. If the logic thread needs to communicate with a window, then it is done like it is described above. But there are certain requests that have no windows associated. For this reason the GUI thread upon startup maintains a window invisible to the end user. This is a blind window that serves as a gateway for certain requests. One of these requests is creation of a GUI object like a normal window or a dialog box. As it has been mentioned

above, all GUI objects must be created in the GUI thread that is running a message loop – an unavoidable prerequisite of any windowing – but this is the logic thread that decides when to create a GUI object. Hence, the logic thread through the blind window asks the graphic user interface thread to create an object and return back a pointer on it to the logic thread. Another request issued through the blind window is to quit the application. When the logic thread has finished calculations it is ready to terminate the program. But the program can not terminate because the GUI thread is in the infinite message loop. In traditional applications the user has to close the main application window manually, thus interrupting the message loop. In the case of the PROBE, the blind window object receives a request for program termination through the second type method, which means posting of a "quit" message. The "quit" message leads to the message loop interruption and termination of the application.

Exceptions and Premature Stop

The multithreaded architecture makes the support of exceptions more difficult. The exceptions are incidents like run-time errors, problems with resource allocation and premature halt by the user. The first two types of exceptions are easy to handle through standard C++ exception support but the premature stop is more difficult. The text-based application is normally interrupted by Ctrl-C that forces the program to terminate immediately. But in the case of GUI, which is placed into another thread the situation is more complicated. Indeed the premature termination exception is coming from the GUI thread – the user presses the "STOP" button (Figure 2-2) – but it is the logic thread that must react on it.

Figure 2-2. A main PROBE window during calculations.

Obviously upon the exception one could terminate the program roughly by killing the logic thread and this would work unless the additional features like COM objects

are implemented (see below in this chapter). Simple killing of the logic thread would lead to unreleased resources and unfinished COM processes. To avoid this, a special object ThreadWatchdog has been designed. The only task of this object performed in its single method (constructor) is checking if the GUI thread is indicating a premature stop exception.

A variable of the type ThreadWatchdog has been added to each non-GUI object of PROBE. The ThreadWatchdog object is a part of a new paradigm as well. With available software development packages (for example Microsoft Visual Studio) it is easy to add a variable to a lot of objects without any manual rewriting of them. The idea to add the variable of the ThreadWatchdog type stems from the fact, that whatever the logic thread is doing, it is after all creation and destruction of its objects. But on the other hand, the C++ compiler automatically invokes constructors of all variables that are members of any objects. Hence the following scenario is taking place:

- GUI thread indicates an exception,
- logic thread is still doing calculations,
- at a certain moment the logic thread creates some object X,
- constructor of the ThreadWatchdog is invoked due to the latter is a part of X,
- constructor of ThreadWatchdog rises a C++ exception,
- the C++ exception is caught in the logic thread,
- capture of the exception leads to the automatic memory release, termination of COM processes, stack unwinding and termination of the logic thread.

The scenario is performed fully automatically, solely by the C++ support. Therefore, the logic thread need not be modified from its text version except addition of the ThreadWatchdog variable, which can be done semi-automatically by the software-developing package.

Additional Features

The Component Object Model (COM) is a very powerful technology of Microsoft Windows. Another name of this technology is Object Linking and Embedding (OLE). This technology generalizes the idea of object-oriented programming where the objects can be written in any language. The objects are also compiled with a corresponding compiler and exist in a form of a binary code, for example .ocx, .dll or .exe file [29]. The objects exert interfaces – the means to communicate with them. The interfaces make COM objects similar to the C++ objects. Microsoft Windows takes care of all underlying processes dealing with allocation of the object binary code on the disc, loading this into the memory, invoking methods etc.

PROBE makes use of the COM by collaborating with Microsoft Excel. The output from the GUI PROBE is not just a text, rather it immediately goes to the Excel sheets, enabling the user directly to order the probes from a company or do any further processing of the probes by means of Excel. The most powerful aspect of COM in this case is that the PROBE does not pay attention on how the Excel sheet files are binary organized. Even more, from version to version of Excel the binary structure of its .xls files may change. Instead, the PROBE invokes Excel through COM, asks for the worksheet object, receives a pointer to its interface and puts the output into the cells using the standard methods very well described in the Help system of Excel. Excel takes care of how to process the data and how to store them on the disk.

A Sight on PROBE

Putting it all together here are the examples of dialogues offered by PROBE. The dialogue shown on Figure 2-3 is popped up during the startup. This dialogue determines the mode of calculations to be performed.

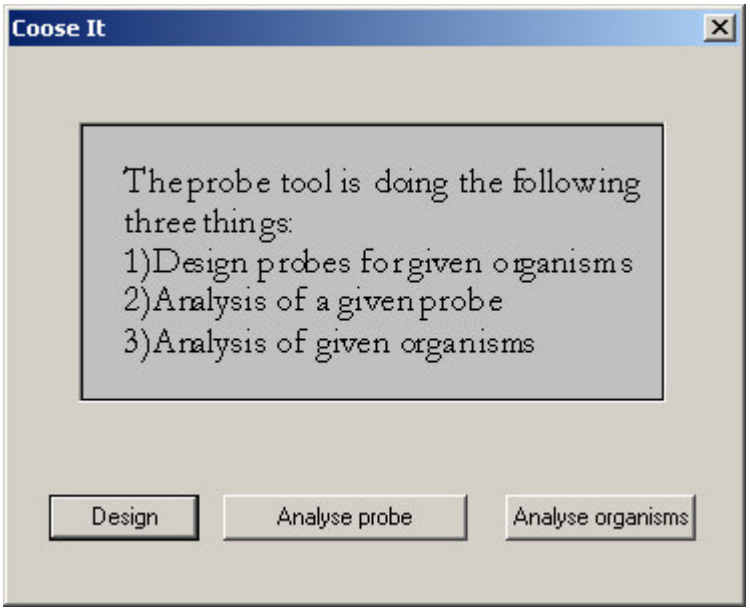


Figure 2-3. A start-up dialogue determining the computation mode of PROBE.

One can see that not only the controls presented on the dialogue window, but also a clear short explanation of the modes. The user would intuitively understand the option "Design" even without reading the article [64] describing the program – the main purpose of the PROBE is indeed the design of oligonucleotide probes. After this dialogue several others appear among which is the one presented on Figure 2-4.

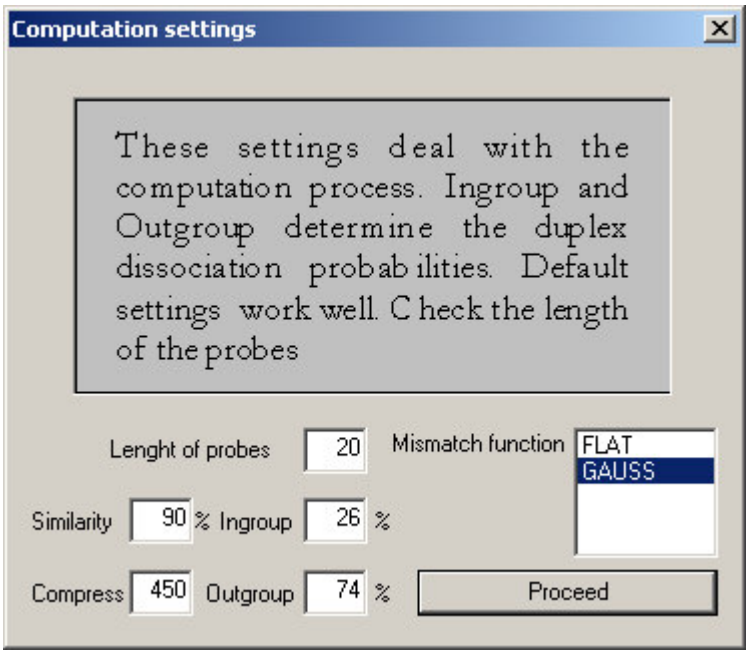
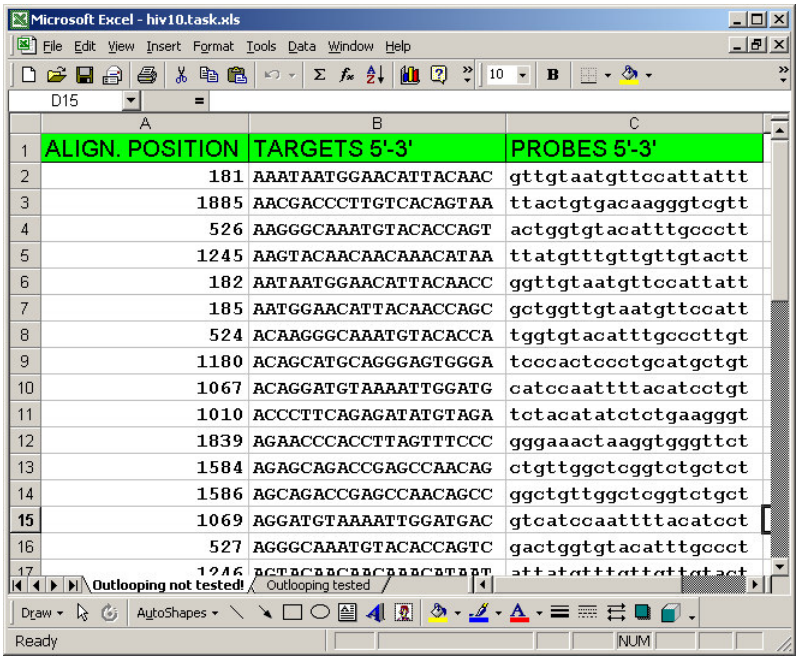


Figure 2-4. Computation settings of PROBE.

This dialogue determines settings of the calculation process. Again a short preamble explains what the settings mean for the computation. The optimal defaults are already provided. The detailed explanation of the settings can be found in the

article [64] or on the web site of the program: http://biochip.genetik.uni-koeln.de/probe. But even if some of the settings are unclear without reading further information, it is possible for the user to use the defaults and perform his or her first calculation. Psychologically it is much more convenient to go through the whole process at least once and dive into the greater details only in case if something wrong happens.

The output of the GUI version of PROBE, as it has been stated above, is written directly into the Excel sheets, allowing the user to make any further processing with the designed probes if necessary. The table has three columns as one can see on Figure 2-5. The leftmost column shows the alignment positions of the 5'-end of targets, with which the corresponding probes would hybridize. The middle column shows sequences of the targets and finally the rightmost column shows the probes themselves. These probes are to be immobilized on the chip.



	A	B	C
1	ALIGN. POSITION	TARGETS 5'-3'	PROBES 5'-3'
2	181	AAATAATGGAACATTACAAC	gttgtaatgttccattattt
3	1885	AACGACCCTTGTCACAGTAA	ttactgtgacaagggtcgtt
4	526	AAGGGCAAATGTACACCAGT	actggtgtacatttgccctt
5	1245	AAGTACAACAACAAACATAA	ttatgtttgttgttgtactt
6	182	AATAATGGAACATTACAACC	ggttgtaatgttccattatt
7	185	AATGGAACATTACAACCAGC	gctggttgtaatgttccatt
8	524	ACAAGGGCAAATGTACACCA	tggtgtacatttgcccttgt
9	1180	ACAGCATGCAGGGAGTGGGA	tcccactccctgcatgctgt
10	1067	ACAGGATGTAAAATTGGATG	catccaattttacatcctgt
11	1010	ACCCTTCAGAGATATGTAGA	tctacatatctctgaagggt
12	1839	AGAACCCACCTTAGTTTCCC	gggaaactaaggtgggttct
13	1584	AGAGCAGACCGAGCCAACAG	ctgttggctcggtctgctct
14	1586	AGCAGACCGAGCCAACAGCC	ggctgttggctcggtctgct
15	1069	AGGATGTAAAATTGGATGAC	gtcatccaattttacatcct
16	527	AGGGCAAATGTACACCAGTC	gactggtgtacatttgccct
17	1246	AGTACAACAACAAACATAAT	attatgtttgttgttgtact

Figure 2-5. An example of the output provided by the GUI version of PROBE.

Conclusion

A new paradigm for obvious graphic user interface applications has been developed. The PROBE has been powered with GUI. The GUI version, like its text ancestor, is not waiting for the user's actions; instead it is guiding the user through the whole process of the probe design. The output is directed into the Excel worksheets enabling easy further processing or ordering the oligonucleotides from the supplier.

Chapter 3 Dissociation kinetics

Introduction

Biochip technology nowadays offers a convenient way for microbial identification and expression profiling. This technology implies employment of solid DNA support with immobilized oligos organized in spots. A sample under consideration is applied onto the biochip and its fluorescent-labeled nucleic acids (e.g. rRNA, mRNA) hybridize with the immobilized oligos. One acquires information from the biochip by analyzing the spot intensities. The spot identification and cross-hybridization are focused on in this chapter.

There have been many attempts to solve the problems of cross-hybridization. Here I propose a new approach based on the kinetics of dissociation of nucleic acid duplexes. The exploitation of kinetics for solving a cross-hybridization problem has already been done, but instead of dissociation, the association process was taken into a consideration [30]. Something similar to the dissociation approach is presented in the work of Drobyshev [31], where the author studied a microarray being washed at rising stringency, but this method is a result of a complex overlap of dissociation kinetics and thermodynamic stability of the duplexes. Moreover, the available techniques do not consider cross-hybridization to occur along with the true hybridization on the same spot, which can be the case especially among the oligonucleotide microarrays. The key point of the new approach being proposed is that the cross-hybridized mismatched duplexes are less stable than the perfect ones and have higher rates of dissociation [32-34]. For example, a TT single mismatched duplex dissociates 120 times faster than the perfect matching one [33]. If one allows the mismatched duplexes to dissociate first and disappear from the signal, then it is possible to calculate the initial value of the signal without wrong duplexes.

The identification of spots is another big issue of microarrays. Ideally, the spots should have round shape with uniform intensity and equal radius. The background is ideally uniformly distributed all over the microarray. In reality, the spots have various shapes (donut, round with wrecked edges, etc.) and the background is not uniform, containing bright portions emerging from the dust and other artifacts. Most of the commercially available packages rely on the manual spot assessment, which is very

tedious and slow. Liew [35] proposes an elegant spot recognition algorithm based on the technique used for restoring of the degraded documents [36]. The algorithm is based on the idea of comparing the gray level of the processed pixel or its smoothed gray level with some local averages in the neighborhoods with a few other neighboring pixels. The algorithm relies on the image information only and does not take into account physical properties of hybridization. The idea of the new method being proposed here is that the duplex can dissociate and, therefore, one can observe this process as a decay of fluorescence on the chip, while dust and any other disturbing signals stay constant or have highly irregular behavior. Thus, by recording images during the dissociation process, one makes use of the natural behavior of the dissociating duplexes and multiplies the amount of information that increases the reliability of spot recognition and elimination of artifacts.

Theoretical Considerations

Signal Preparation

Based on the phenomenon that the rate of dissociation depends on the matching quality, the following system of equations (Equation 3-1) describes the time behavior of the signal, where S – fluorescent signal, P – quantity of perfect duplex at the spot per its area unit, U_i - quantity of any (i-th) imperfect duplex at the spot per its area unit, k_p k_{Ui} – corresponding rate constants.

$$\begin{cases} \frac{dP}{dt} = -k_p P; \\ \frac{dU_i}{dt} = -k_{Ui} U_i; \\ \dots \\ S = P + \sum_i U_i; \end{cases}$$

Equation 3-1. Dissociation processes in the presence of cross hybridisation. The letters *P* and *U* designate concentrations of perfectly matching and mismatched duplexes respectively. *S* is a full signal comprised from the contribution of perfectly matching and mismatched duplexes. *k* is a dissociation rate constant.

After rearranging of the system one has to solve it with respect to the rate constant k_p:

$$\frac{dS}{dt} = -k_p P - \sum_i (k_{Ui} U_i);$$

$$k_p = \left(-\frac{dS}{dt} - \sum_i (k_{Ui} U_i) \right) \Big/ P.$$

If enough time is elapsed, the imperfect duplexes have dissociated and washed out, so $U_i \rightarrow 0$ and therefore $S \rightarrow P$. Then the following is valid:

$$k_p = \left(-\frac{dS}{dt} \right) \Big/ S.$$

Equation 3-2. Solution of the problem in terms of the dissociation constant. The dissociation constant is time independent.

This ratio can be easily recorded. The dissociation curve at each pixel of the chip image must be obtained first and then differentiated along its course and derivatives must be divided to the value of the signal where the derivative has been obtained. According to this, we can expect in the experiment the graph depicted in the Figure 3-1 (simulated data).

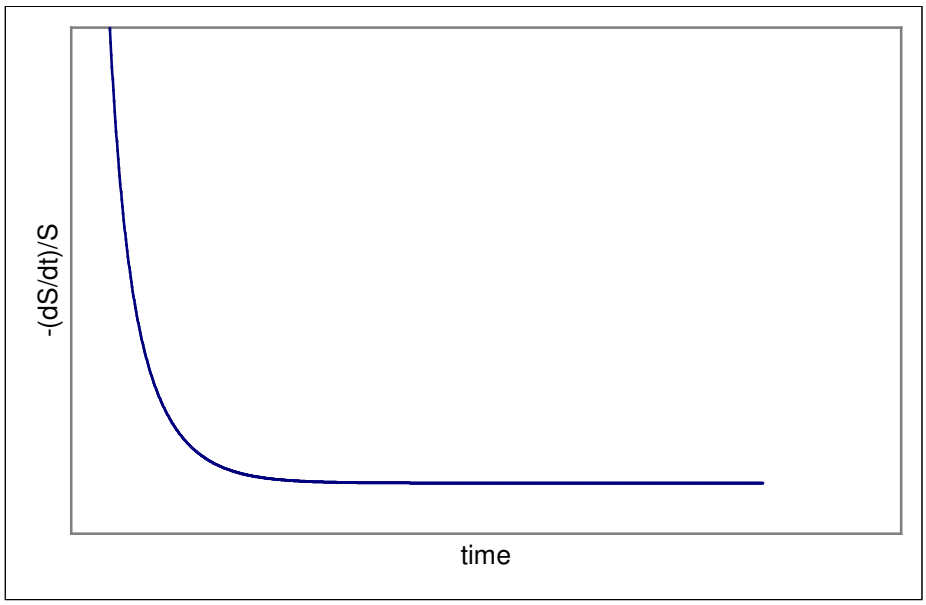


Figure 3-1. Simulated situation of perfectly matching duplex and five mismatched duplexes. After sufficient time has elapsed, a horizontal line represents k_p.

We can expect, that the graph has initial nonlinear and linear parts. After some time the curve approaches a horizontal line. The vertical ordinate of the line corresponds to k_p. Now it is easy to find the concentration of the perfect duplex at the beginning of washing according to the Equation 3-3:

$$P_0 = S_{th} e^{k_p t_h}$$

Equation 3-3. Computation of the initial duplex concentration.

where S_{th} is a value of the signal at time t_h on the horizontal part of the above curve. By this calculation one avoids the signal from cross-hybridized oligos. If one fails to observe the horizontal line, then the spot is unreliable and this is the indication to discard such a spot from the analysis.

Spot Determination and Quantification

In fact, there is no need of explicit spot determination. As it has been described above, the dissociation curves obtained for each pixel of the chip provide the dissociation constant and an initial value of the unbiased signal. Having these quantities, the chip must be represented as two images in terms of the constants and initial values. Dust and empty area of the chip will have zeroes for both quantities and the spots will have certain values. Then it is easy to analyze the image with any software that is able to determine spot locations.

Ranking

To avoid dependency on the experimental conditions, the values of the rate constants are not of great importance at the very beginning of the analysis. Instead, one should rank the constants having a control constant as a minimum. Such a control could be a duplex with much higher melting temperature (i.e. stability) to ensure the slowest possible rate of fluorescence decrease (produced mainly by photobleaching). Having the smallest possible constant, the ranking procedure will help to differentiate specific and unspecific matching (already without cross-hybridization). Moreover the ranking helps to find out the resolution of the whole method: if the range of the constants from the control one to the highest one is very small, then the resolution is very poor and one should increase their difference by changing the experimental conditions.

Hybridization and dissociation

The key point of the experiment is that hybridization and dissociation are carried out in the same buffer and at the same temperature. The driving force for the dissociation is absence of nucleic acids in the washing solution (equal to the

hybridization buffer). The entropy then drives the duplex dissociation. In fact, the isothermal isobaric process happens spontaneously when and only when the differential of the Gibbs energy is negative at all conditions being fixed, but deliberating a break apart of an infinitesimal amount of duplexes:

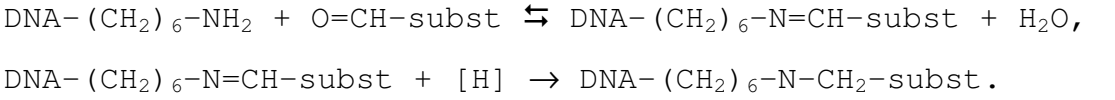
$$dG = dH - TdS,$$

where dH is the heat necessary to break up the duplex, T – temperature, dS – the entropy gain after breaking up. In the process being considered, the spontaneity is the case: TdS is always larger than dH if the excess of the washing solution contains no DNA.

Method Establishment

Super Aldehyde slides

First of all it was necessary to be able to produce oligonucleotide microarrays and make sure that DNA is indeed immobilized. Super Aldehyde microarray substrates were the first choice to start with. The immobilization procedure employs the reaction of the C6 amino modified oligonucleotide with the aldehyde group stemming from the substrate's surface. At the first reversible stage the Schiff base is formed, which is then on the second stage irreversibly reduced by sodium borohydride (Scheme 3-1).



Scheme 3-1 Formation of the Schiff base and subsequent reduction (with $NaBH_4$).

Microarrays were created according to the scheme mentioned above (see Materials and Methods) with various concentrations of the oligonucleotide P1 (see Table 3-1) in the spotting mixture: 50.0, 37.5, 25.0, 12.5 and 0.0 μM. To determine the quality of immobilization, the SYBR Green staining was employed. This technique is well established for cDNA microarrays [61]. Strikingly, the staining displays maximal signal where no DNA is located (Figure 3-2).

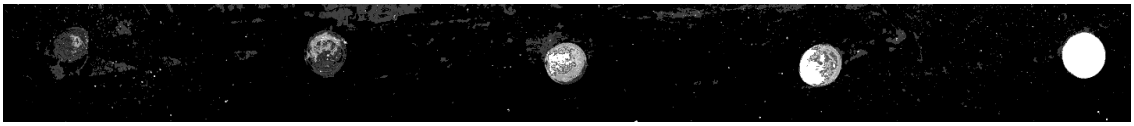


Figure 3-2. SYBR Green staining. Amount of the oligonucleotide decreases from left to right. Staining shows directly the opposite.

Because of the unavailability of the information from Telechem, a possible "phenomenological" explanation could be proposed that the Micro Spotting Solution (Telechem) contains some compound "X" that binds to the glass slide and this compound is readily stainable with the dye. When there is DNA in the solution, the DNA binds instead of the "X". But DNA in this case is a single stranded 20 nt oligonucleotide that is stained very weakly. By the time of the ongoing experiments, another version of the spotting solution – the Micro Spotting Solution Plus – became available. Staining experiments with this product displayed no signal at all, supporting the theory about the "X" compound. The other means to stain the microarrays with SYTO11-16, Panomer 9, OliGreen, according to either their cognate protocols or the one similar to SYBR Green, generally failed. Either the signal was absent, or the spots without DNA were stained, similarly to the already described effect.

A common microscopic slide was used as a control and displayed no signal regardless whether there was or was no DNA in the spots. Another control was the hybridization with 5'Cy5-K1 – a complementary towards P1 Cy5 labeled oligonucleotide (see Table 3-1) at 42 ^{o}C for 1h. In this case a strong signal was observable exactly at the spots containing DNA.

The failed attempts to control the immobilization lead to the necessity to make multiple spots in order to statistically overcome the uncertainty of immobilization. The differences in the amounts of the probe of a certain type at several spots will contribute to the standard deviation of the mean signal read from them, provided that the solution containing complementary DNA that is applied onto the chip for hybridization, has homogeneous distribution of the polynucleotides. The latter is normally the case if during hybridization the agitation is used.

Preliminary dissociation experiment

Having established the microarrays, a preliminary experiment concerning the possibility of dissociation was carried out with P2 immobilized and 5'Cy5-K2

complementary oligonucleotides. Hybridization was set up at 50 ^{o}C for 4 h. The dissociation was performed as described in the Materials and Methods section. During the dissociation the microarrays were incubated in the washing buffer for 0, 1, 2, 3, 8, 15 min. Figure 3-3 shows the results. It is obvious, that there is a prominent fade of the signal indicating certain kinetics.

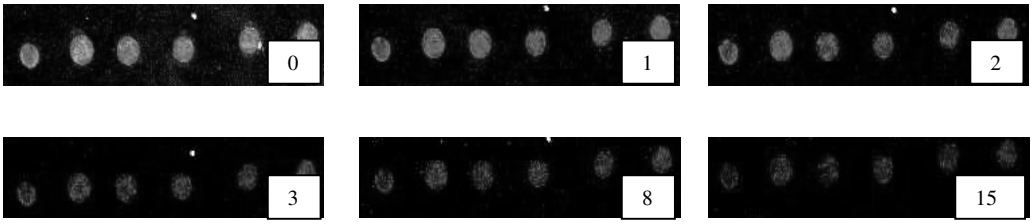
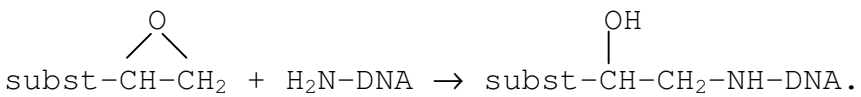


Figure 3-3. Series of images scanned while washing. Upper left image corresponds to zero time, then 1, 2, 3, 8, 15 min. further right and down.

Accurate kinetic analysis is not possible without real-time intensity measurements during dissociation. The scanner available in the lab was not possible to modify for such measurements. An ideal alternative was a Leica laser confocal microscope enabling to install on the object table a washing chamber (described below) and perform real time scanning and image recording. When the microarray, having the P2 immobilized probe and hybridized 5'Cy5-K2 oligo, was observed by laser scanning, a very poor signal came up. The reason was either the poor sensitivity of the microscope or in general weak immobilization capacity of the Super Aldehyde slides. Later on, the latter problem turned out to be the case.

Epoxy slides

Among many available various microarray products, one offers a particular promising immobilization chemistry. Scheme 3-2 below shows that the immobilization occurs via a single irreversible reaction between a very reactive epoxy group and amino group.



Scheme 3-2 Immobilization via the reaction with an epoxy group.

Although the reaction looks very simple, it requires special conditions, namely controlled temperature and humidity. The temperature can be easily adjusted, but humidity tends to decrease when the temperature rises. To achieve the goal, numerous different ways were explored. The best and the most effective approach (at the same

time the simplest one) is using a glass filled up with water only at the bottom and closed with several layers of filter paper (Figure 3-4).

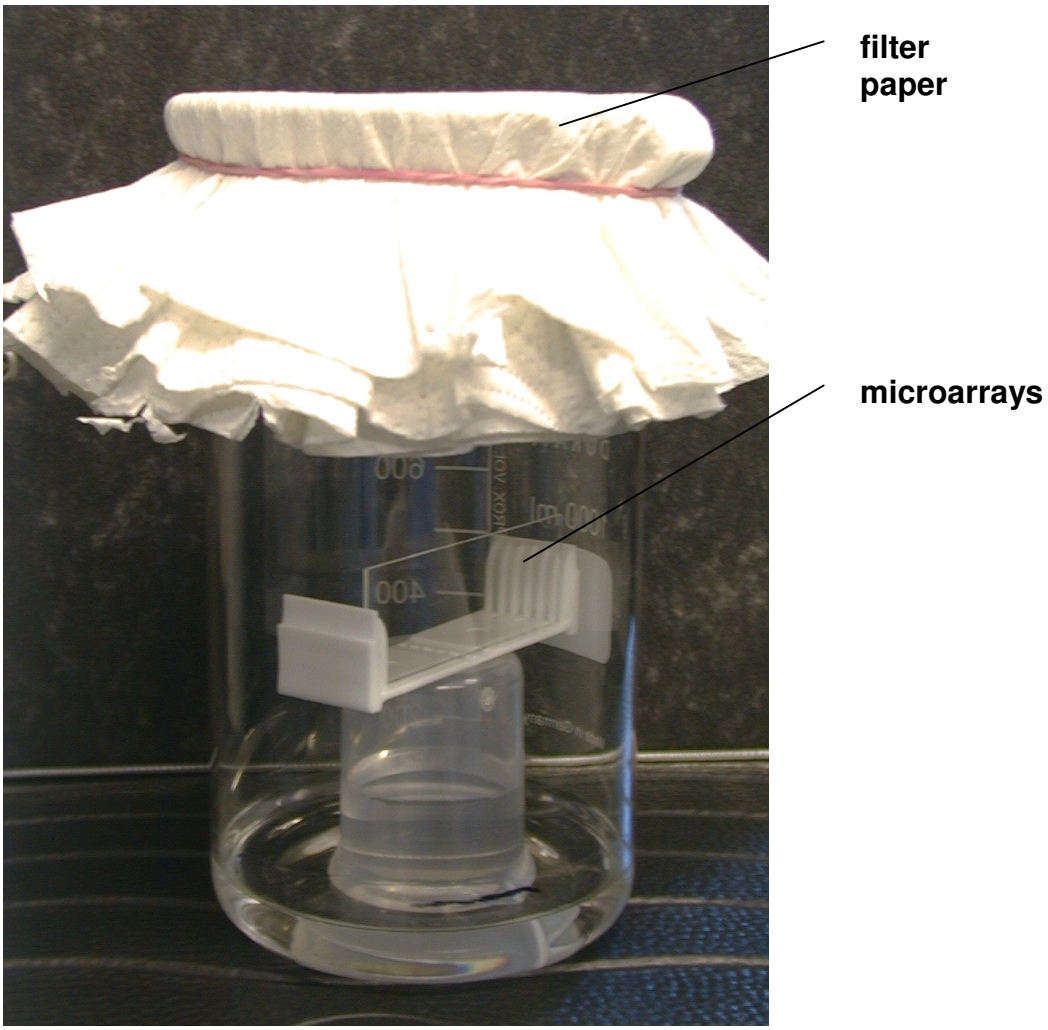


Figure 3-4. Chamber for humidity control during immobilization on the MWG epoxy slides.

The microarrays are placed on a certain elevation above the water level. The glass with microarrays is then placed into the heating oven. The physical principle behind the device is the following: there is a certain profile of the humidity along the height of the glass emerging due to the difference between the humidity right above the water level (100%) and right beneath the filter paper (equal to the humidity in the oven). The profile stays unchanged when a dynamic equilibrium between the amount of vaporized water and water leaving the opening of the glass through the filter paper is established. Thus, at a given height plane perpendicular to the axis of the glass the humidity stays constant. By variation of the opening permeability (by altering the number of the filter sheets) one can establish any desired humidity at the position, where the microarrays are placed. The Materials and Methods section describes the immobilization protocol.

Another important point of the microarray fabrication is the ambient humidity during spotting. It was already published [37] and from the personal experience determined that the optimal humidity must be 55-60%. If the humidity is too low, the needles of the robot, delivering the oligonucleotides to the slide, dry before they reach the slide surface and therefore the transfer does not occur. Setting up the desired humidity posed therefore a challenge. To solve the problem numerous attempts were made. Especially the construction of the spotting robot did not allow to set up the humidity right in the spotting area due to the forced air exchange between the inner camber and environment. Therefore it was necessary to change the humidity in the whole room. A conventional room humidifier (Burg, Germany) at its outmost power could increase the humidity maximally for 5%, taking into account the ambient humidity at the time of the experiments equal to 30%. The solution was a DIY humidifier made of two 8 kW water boilers having a PC fan for vapor spreading. One of the boilers worked constantly, another was automatically switched on and off by a DIY hydrostat (Figure 3-5) tuned to 60% humidity. The DIY hydrostat was constructed from the Feuchtigkeitsschalter kit available from Conrad Electronik, Germany.

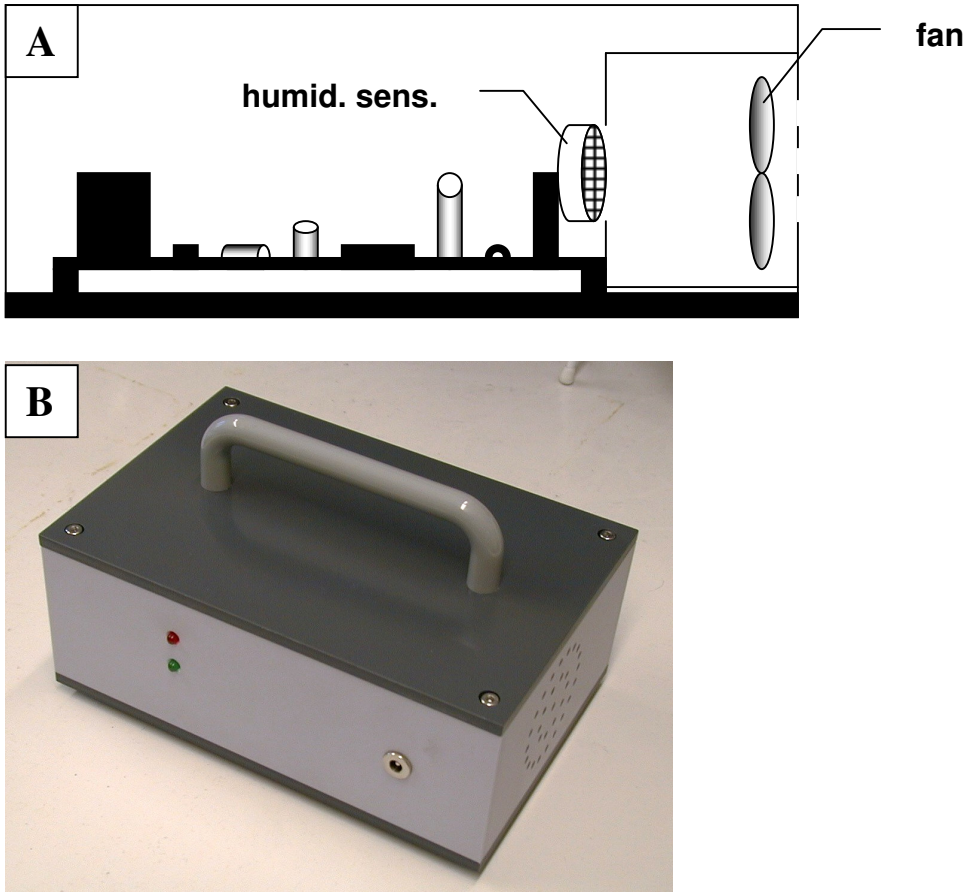


Figure 3-5. A) **Scheme of the hydrostat. A fan drives the air onto the humidity sensor.** B) **Outlook on the device.**

The microarrays based on the MWG epoxy slides displayed a staggering difference in the immobilization extent. It became possible to observe them under the confocal microscope. Figure 3-6 shows the images of the microarray, having P1 immobilized and 5'Cy3-K1 hybridized oligonucleotides, recorded with the confocal microscope and the GSI scanner.

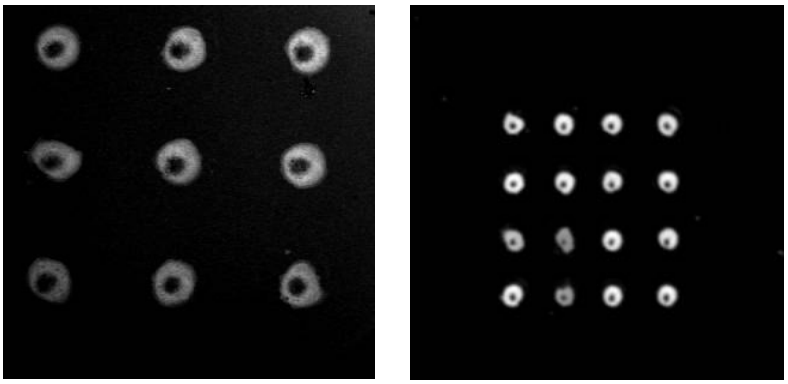


Figure 3-6. A microarray based on MWG epoxy slides. The views under the confocal microscope (left) and GSI scanner (right).

Dissociation setup

A constant buffer flow of a controlled temperature through the dissociation chamber was established by the kinetic setup depicted on Figure 3-7 along with the scheme of the working principle.

The reservoir equipped with a magnetic stirrer contains a washing buffer, the buffer is heated up by an immersion heater, which is in turn controlled by a pre-calibrated switch. The switch turns ON when the temperature sinks lower then the desired value and turns OFF otherwise. The temperature is measured in the dissociation chamber (see below) and a feedback loop is established between the chamber and the switch. The switch was constructed from the Temperaturschalter kit available from Conrad Electronik, Germany and fine-tuned with the MeasureSuit [38]. The buffer flows from the reservoir through the setup due to the atmospheric pressure. Behind the chamber the buffer is collected and returned back to the reservoir by a peristaltic pump.

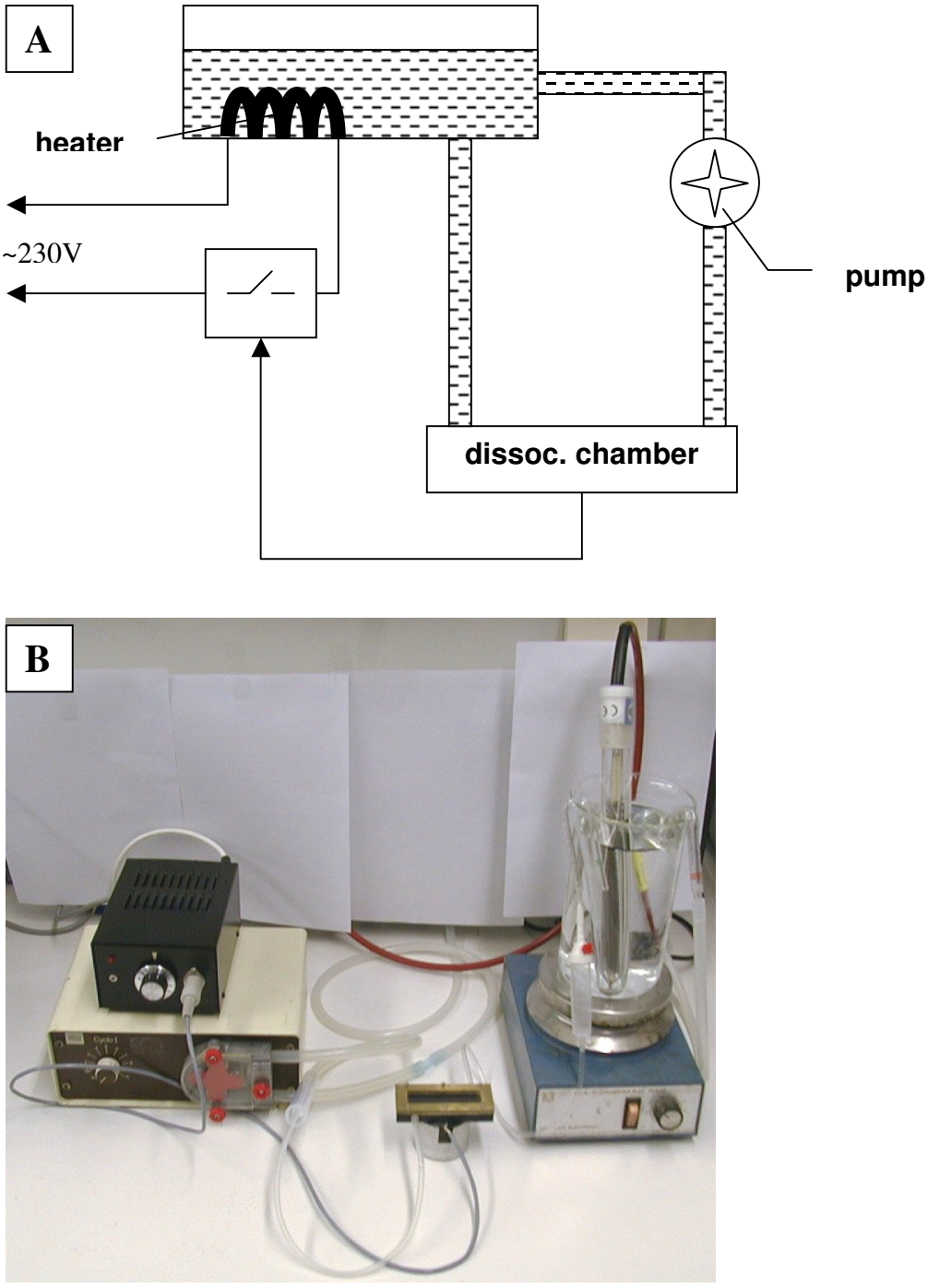


Figure 3-7. Dissociation setup. A) **scheme of the setup.** B) **photograph of the implemented setup.**

The dissociation chamber for recording of real-time kinetics was constructed. Figure 3-8 shows the chamber from the top view along with the diagram illustrating its working principle.

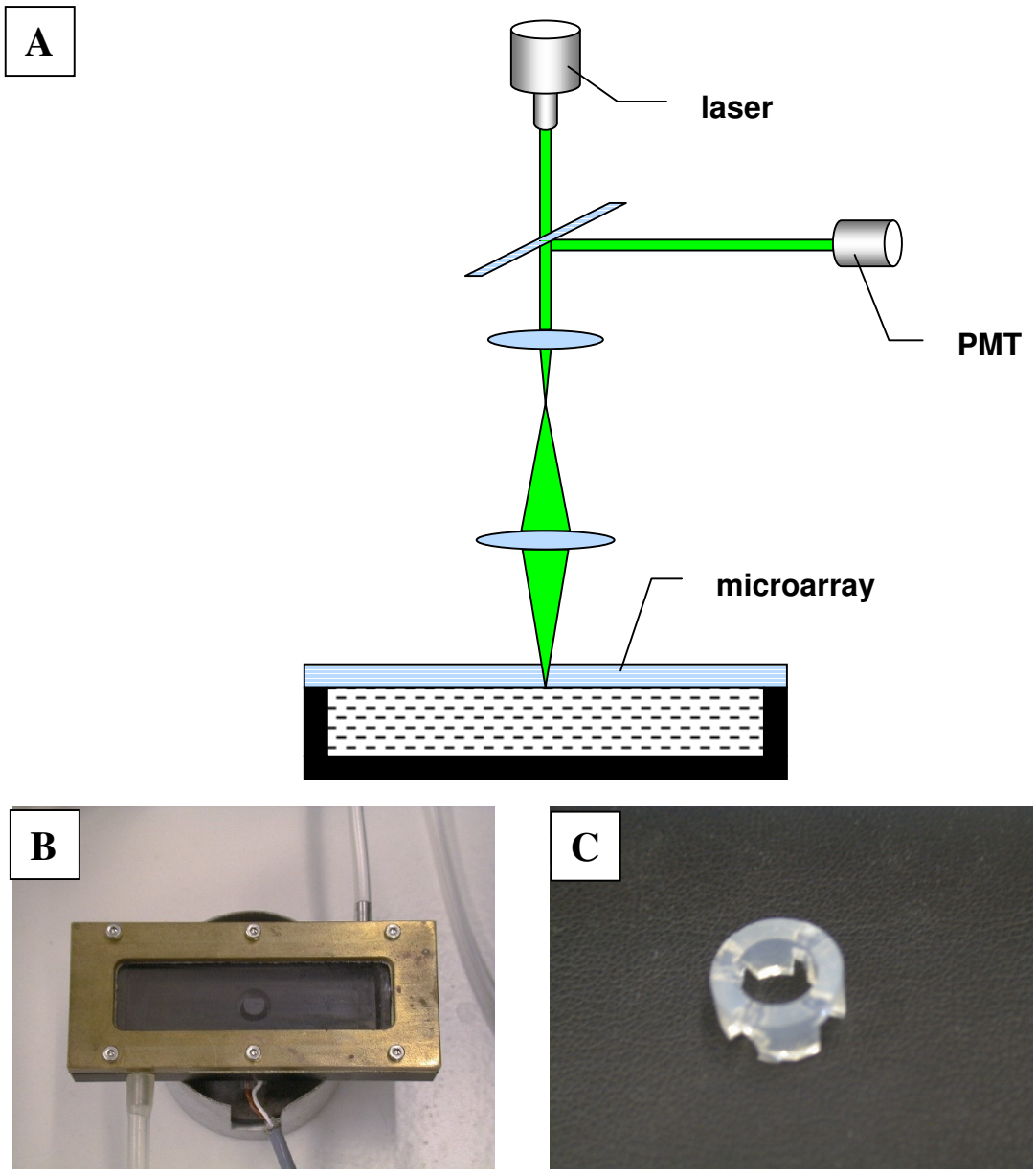


Figure 3-8. A) **Working principle of the dissociation chamber.** B) **photograph of the dissociation chamber.** C) **debubbler to be inserted in the chamber.**

The chamber is a flat prolonged box, with the microarray slide as a top plane. The size and form of the chamber enables it to be placed onto the object table of the confocal microscope. The microarray surface containing spots is oriented inwards the chamber. The microarray itself is fixed with a metal frame and 6 screws. The junction between the slide and the rest of the chamber is made watertight with a rubber layer. The chamber has an inlet and outlet for a liquid. In the middle of the bottom plane of the chamber a thermo-resistor sensor is placed, which wires are lead outside of the chamber for the external control.

The experiments with the chamber revealed an intensive formation of bubbles obscuring the scanning area. Many different tricks to avoid the bubbles were tried. Among them there were the use of a blood transfusion system with its dedicated debubbler, vacuum degasation of the buffer and pre-heating of the setup to the working temperature in order to desorb the air from the surfaces. All of these techniques were generally unsuccessful due to the unavoidably open buffer reservoir catching the air from the surrounding. A final solution appeared to be a piece of a silicon hose surrounding the spots of the microarray. The hose with its one end is tightly attached to the microarray, with another to the bottom of the chamber. The walls at the end of the hose attached to the bottom of the chamber have "windows" for the buffer flow. When the bubbles are formed, they are moving with the buffer flow and since they are lighter than the buffer, they slide over the microarray surface. When the bubbles encounter the hose, they can not enter the windows of the hose – the bubbles would have to sink otherwise (see Figure 3-8). Hence the bubbles are forced by the liquid current to bypass the hose and therefore the area being scanned.

Indirect Labeling

The dissociation kinetics studies are intended to be carried out with native ribosomal RNA and therefore a labeling procedure must be established. There are numerous labeling procedures. The most popular methods for nucleic acid labeling are currently based on enzymatic procedures such as those involving reverse transcriptases [40-44], terminal transferases [45, 46], kinases [46], random priming [47, 48], or PCR [49-57]. Most of these protocols also demand careful nucleic acid purification, separate sample fragmentation procedures (which considerably improve the specificity of hybridization), and a final precipitation or gel filtration step to eliminate excess label. As a result, sample isolation and fractionation steps usually precede separate labeling-fragmentation-purification routines. Recently developed chemical labeling methods also require a considerable time to perform (more than 3 h) [58, 59]. Another very fast labeling procedure taking 20-30 min was developed by Bavykin, et. al. [60]. This procedure is based on the oxidative cleavage of RNA with formation of active species reacting with amino derived fluorescent labels. In order to simplify labeling of RNA and at the same time increase accuracy of quantification by having a single label per a single RNA molecule I have developed an indirect labeling method. The principle of the method is shown on the Figure 3-9.

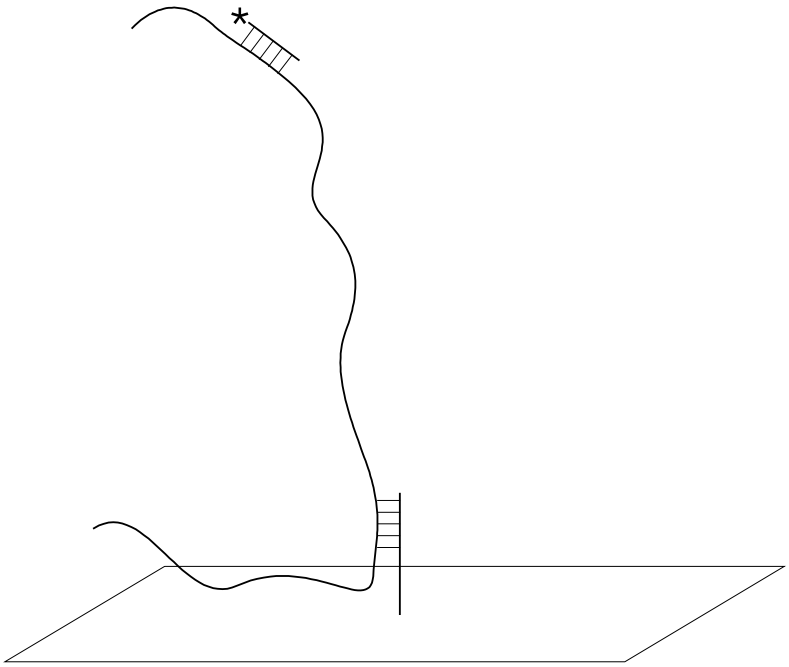


Figure 3-9. Indirect labeling. On the one hand RNA hybridizes with an oligonucleotide of the microarray, on the other hand it hybridizes with the "marking oligo", labeled with a fluorescent dye (indicated by an asterisk).

A labeled oligonucleotide (marked by asterisk) – the "marking oligo" – hybridizes with a conserved region of RNA, while all the variable regions are available for hybridization with specific oligonucleotides of a microarray. This method is actually a refinement on already published technique [39]. The authors insist on the proximity of the "marking oligo" to the region that is detected by the specific oligo as far as 3 bases apart. In their case the detection limit was 50-500 pM. The method used in this work does not require proximity and the determined detection limit of the method is not worse than 300 pM of rRNA (60 pg/µl). The detection limit was determined using a standard microarray format with the chip containing the P2 immobilized oligonucleotide. The hybridization was performed as described in the Materials and Methods section with rRNA of Algae01 species (obtained by *in vitro* transcription of the respective clone) at different dilutions decreasing the concentration of rRNA. The smallest concentration that still produced well recognizable signal above the background level was considered as a detection limit.

Software

Recording of the real-time kinetics implies regular scanning of the microarray. The Leica confocal microscope is able to perform scans one after another spaced with a desired time interval. As the result one receives a series of TIFF files named with sequential numbers. To discover the dissociation curve a computer program was created. The program reads the series of tiff files, converts their names into the corresponding time points and looks after the intensity of the same pixel on all images. The program has options to perform averaging within a square patch or to integrate the spot intensities within the square. The output of the program is a text file of time-intensity pairs. The program was designed in two versions for a single computer and a cluster (or a multiprocessor computer). The program makes use of the TIFF processing library available elsewhere [63].

Dissociation Experiments

The first experiment was carried out with the chip having the P1 immobilized oligo and hybridized with the complementary one 5'Cy3-K1. The dissociation curve was recorded at 45 ^{o}C, washing with 5xSSC. The curve is shown on Figure 3-10A, the intensity of the pixels was averaged within 16x16 rectangle. The time course of fluorescence has an initial steep part and slow rest. Moreover, the slow part of the curve has a wave-like quality. This wave-like behavior became much more prominent after the numeric derivation of the curve, that made it impossible to analyze it accurately (data not shown). Several things had to be considered. Does the observed fluorescence change indeed correspond to the dissociation or the fluorescence decay happens due to the other reasons? Why does the dissociation curve have two such distinct decrease rates while the single duplex is supposed to dissociate uniformly? What is the wave component of the curve, obviously having nothing to do with the dissociation itself and dramatically disturbing the analysis?

The stability of the instruments was the first issue to be checked. Instead of the spots on the microarray a piece of aluminum foil was glued onto the normal microscopic slide and the same series of images was recorded. In this case not the fluorescence, but the reflected light was captured by the PMT and quantified. The Figure 3-10B shows the time course of the measured signal. Strikingly, there is a prominent decay, that can be attributed either to the laser intensity drift or shifting

sensitivity of the PMT. On the other hand one can see that after about an hour the decay is getting quite slow. In fact, the same experiment was performed but after running the microscope idle for 1.5 hours. The signal intensity then decays much slower, but the wave component is somewhat increased (data not shown). This tells us that the pre-heating stabilizes the instrument, either it does with the laser or PMT or both of them.

There is an indication in the literature, that the PMTs tend to have uneven sensitivity during time. The sensitivity has overlapped long and short periods of alterations of unknown origin [62]. Apparently, the reference is quite old, while nothing new is available, and the technology of PMT fabrication might have changed, it was necessary to examine the issue of the PMT sensitivity in greater details. An assay of the PMT sensitivity was carried out by recording of the intensity of the constant light source. For this purpose a piece of black cardboard with a needle sized aperture was glued onto a conventional microscopic slide and lit from the beneath by a transparent-light source of the confocal microscope. The transparent-light source can be considered stable because this is a simple incandescent lamp powered from the stabilized power supply and has no obvious or documented reasons to be unstable. Before scanning, the microscope was pre-heated by running idle for 1.5 hours as described above. The scans were processed with the software described above using the option to integrate the pixels within the square covering most of the light spot. In contrast to the single-pixel tracing, this option spectacularly reduces sporadic glitches of intensity. Figure 3-10C shows the time course of the PMT sensitivity. The waves mentioned above are prominent along with the slow general decay of sensitivity. The period of the waves is about 20 min that is much larger than the mains AC voltage frequency, thus making them most likely to be coming from the PMT rather than from the possible light source variations. From this moment it becomes clear that all the measurements performed with the PMT (and of course with the confocal microscope) must be corrected by having the constant light source as a reference. The correction factor is then:

$$f_t = \frac{R_t}{R_0},$$

where R_t is the reference signal at the time point t, and R_0 is the reference signal of the very first measurement when t= 0. To obtain correct measurements, this correction factor must be multiplied with the measurement to be corrected.

Another aspect of the fluorescence measurements that had to be taken into account is possible photobleaching of the fluorescent dye. The assay of the fluorescence steadiness was performed with the chip having the P1 immobilized oligo and hybridized with the complementary 5'Cy3-K1. The microscope was preheated for 1.5 hours and as a reference the transparent-light source was used. The corrected time course is displayed on the Figure 3-10D. Apparently during 2 hours there is no significant photobleaching of Cy3, which means that the kinetic assay with this dye does not need to be corrected for the photobleaching. In contrast, the same experiment with 5'Cy5-K1 shown on Figure 3-10E shows a clear photobleaching time course.

The final experiment based on the previously collected knowledge was carried out with in vitro synthesized rRNA. The microarray contained P2 immobilized oligonucleotide while rRNA had a complementary stretch towards it. The RNA was hybridized at 45 ^{o}C for 15 h and washed in the kinetic setup with the mixture of 5xSSC and 0.1% SDS. The microscope was initially pre-conditioned by running idle for 1.5 h. As s reference the conventional microscopic adjustable light source was employed. The light was directed on a piece of aluminum foil situated on the outer surface of the microarray in a vicinity to the spots being scanned. Therefore the images displayed spots and a part lit by the reference light. The initially obtained curve was corrected with the reference. The time course of fluorescence is shown on Figure 3-10F. There is a large glitch disrupting the otherwise normal exponential decay. The glitch occurred due to the slide breakdown.

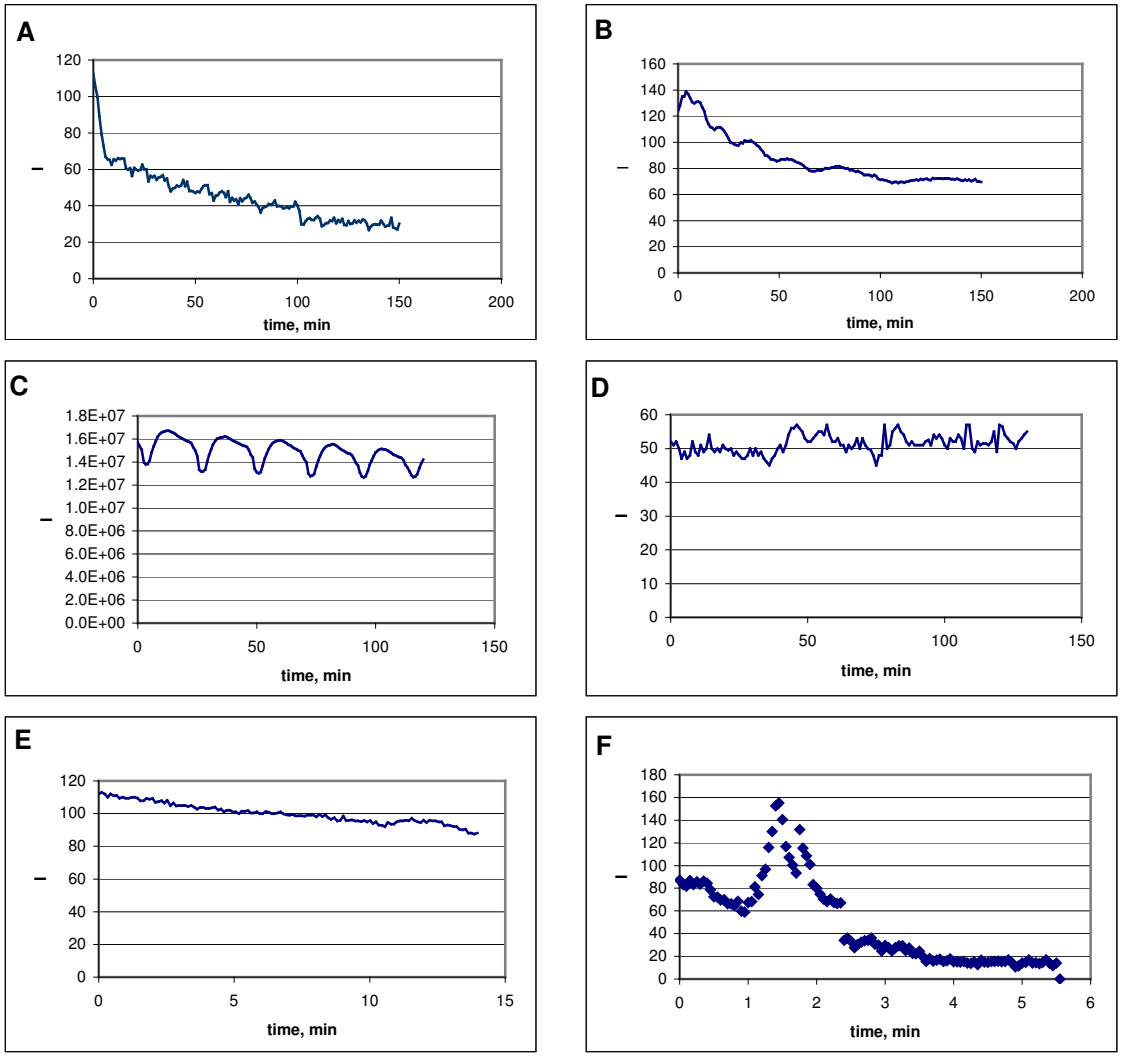


Figure 3-10. Dissociation records. A) **uncorrected dissociation curve, oligonucleotide-oligonucleotide diplex.** B) **recording the time course of the intensity of the reflected laser light.** C) **time-course of the PMT sensitivity with a constant reference light.** D) **photobleaching of Cy3, corrected curve.** E) **photobleaching of Cy5, corrected curve.** F) **dissoiation of the rRNA-oligonucleotide duplex, corrected curve.**

Conclusion

Making the conclusion it should be noticed that successful studies of dissociation kinetics in order to perform match-mismatch discrimination are possible with a confocal microscope. The nature of the detector (PMT) requires its pre-conditioning and obligatory use of a reference light. On the other hand the scanning must be accompanied with a continuous buffer flow washing the microarray. The temperature of the buffer in the chamber containing the microarray must be controlled automatically. For all of these conditions to meet, there was a great difficulty

concerning available equipment and manpower. Therefore these experiments have been suspended for the future.

Materials and Methods

Immobilization of oligonucleotides onto the Super Aldehyde substrate (ArrayIT, USA) was performed by manual spotting. The spotting mixture was prepared by mixing of equal volumes of the oligonucleotide aqueous solution and 2x Micro Spotting Solution (Telechem, USA). Oligonucleotides are listed in the Table 3-1, produced by Eurogentec, Belgium. The spotted slides were dried in an silica-gel exsiccator for 20h. After this, the slides were rinsed in 0.2% SDS for 2 min. and in Millipore water for 2 min. The slides were put into boiling Millipore water for 3 min, dried by centrifuging at 500g, 1 min. in Falcon tubes and reduced for 5 min. by 1.5 g $NaBH_4$ in 450 ml PBS mixed with 135 ml ethanol. After the reduction, the slides were rinsed twice with 0.2% SDS for 1 min. and put for 5 s. into boiling Millipore water. Finally the slides were dried by centrifuging as before.

Staining with SYBR Green was performed by 1x SYBR Green (Molecular Probes, Netherlands) in TBE for 4 min. After that, the chips were rinsed 3 times with TBE and dried by centrifuging at 500g, 2 min. To remove SYBR Green the chips were treated by the solution containing SDS 0.1%, EDTA 1mM, TrisHCl 10mM, pH 7.5 for 1 h, then rinsed with Millipore water and dried by centrifuging.

Hybridization experiments were performed in the Hybridization Station (Gene TAC, USA) at various temperatures for various times in the hybridization solution, containing 5xSSC, 0.1%SDS, 0.2 mg/ml BSA (New England Biolabs) and 10 μM of the complementary probe. Washing was carried out 10 times by a mixture containing 1xSSC and 0.1% SDS at 19 ^{o}C, 1 min. flow, 30 s. hold. Finally the chips were dried with compressed air for 2 min. Scanning was performed with the GSI (Gen TAC) scanner with appropriate laser and filter settings.

Hybridization involving an indirect labeling was essentially carried out as described in the previous paragraph, but the hybridization solution additionally contained 5 μM of marking oligo `5′-Cy3-CTC-CTT-GGT-CCG-TGT-TTC-AAG-ACG-G-3′` and 10 mM Ribonucleoside Vanadyl Complex (New England Biolabs).

For the complimentary material an *in vitro* synthesized rRNA was employed. Final concentration of the rRNA was 1 μM.

Cloning and *in vitro* RNA synthesis were carried out as described in the Materials and Methods section of the Chapter 4.

Preliminary dissociation experiments were performed in the 250ml glass immersed into the thermostatic bath at 50 ^{o}C. The glass contained 1xSSC, 0.1%SDS as a washing buffer. Microarrays were completely immersed into the washing buffer and incubated for various times.

The microarrays based on the MWG Epoxy slides (MWG, USA) were created by spotting of the 5' C6-amino modified oligonucleotides (Metabion, Germany), delivered onto the slides by the Gene TAC spotting robot (Genomic Solutions) at the ambient humidity 60%. The immobilization of the modified oligonucleotides was performed at 42 ^{o}C, 50% humidity for 8-12 hours. The necessary humidity was established in a 1L glass of 180 mm height containing 100 ml of water. The microarray slides were put at the 40 mm elevation above the water level. The glass was tightly sealed with several sheets of filter paper. The number of sheets was determined empirically using a conventional hygrometer so that the elevation level, at which the slides were placed, reached 50% humidity.

Recording of images for real-time kinetics was performed on the Leica Confocal Microscope under control of its cognate software. Kinetic data were extracted from the series by a self-designed program. Further analysis of the data was performed within Microsoft Excel.

Tables

Table 3-1. Set of oligonucleotides for the method establishment.

ID	Sequence 5'-3'
P1	AmineC6-GGT-TTT-TTA-ACC-CGC-AAA-CT
K1	AGT-TTG-CGG-GTT-AAA-AAA-CC
P2	AmineC6-GAT-TGT-GCA-ATA-CTC-CAA-CC
K2	GGT-TGG-AGT-ATT-GCA-CAA-TC

Chapter 4 Experimental evaluation of the PROBE

Introduction

Chapter 1 describes a new probe design algorithm and program [64]. In contrast to existing solutions, the algorithm allows working with datasets that need not to be carefully aligned and that takes into account the position of mismatches along the recognition sequence. In that work [64] we have proposed an *ad hoc* stability function taking into account the positions of mismatches. This chapter is devoted to the experimental test of the validity of this function. The most important point that is examined here, is the dependency of the DNA duplex stability on the position of mismatches, because this was the key point of the new strategy of the probe design.

Since the end of 2001, when the work on the algorithm was published, a certain progress in the probe design as well as in the knowledge of DNA duplex stability has taken place. Various probe design programs and approaches became available [65-69]. They rely either only on the alignment [66] during finding the signature sequences, or on the nearest neighbor prediction of the stability of DNA duplexes [65, 67, 68] ignoring positions of mismatches, while taking into account possible secondary structure formation. The nearest neighbor model was developed for the duplexes in the solution (not formed on the surface of the microarray) and does not care about the position of mismatches [70-74]. The approach proposed by OV. Matveeva [69] uses a complex thermodynamic consideration for refining a probe selection produced by any program taking into account a compound effect of the secondary structure, probe-to-probe interaction and a base composition of the nucleic acids participating in the microarray experiment. Again this approach relies on the nearest neighbor model developed for solutions. Concerning the influence of the position of a mismatch on the duplex stability there is a general rule that a mismatch near or at the terminus of a short duplex is less destabilizing than an internal mismatch [75]. The group of Stahl completed recently a thorough study in this field [76, 77]. First, it is essential to point out, that the format of the system the group of Stahl is working with, is not exactly compatible with the conditions, for which the dissociation model (being tested here) has been developed in our work. In fact, our model in the equilibrium conditions (constant temperature, and a salt concentration) predicts the concentrations of perfectly matching and mismatched duplexes on the

surface of a microaray where the probes are attached to the glass, while the system of Stahl's group employs nonequilibrium melting at rising temperature with washing at changed salt concentration from the microarray having gel pads as spots. This means, that their results can be considered as hints supporting our model. Their work from 2002 [76] shows a little difference in the behavior of the duplexes mismatched at the terminal positions (from 1 to 4), while the later study [77] shows at some point very strong differences in the remaining duplex concentrations depending on the position of a single mismatch during the course of nonequilibrium melting and washing. Another indication of the positional dependency of the stability is provided for the short octamers by surface plasmon resonance [78], where it was shown that the intermediate mismatches destabilize the duplex at greater extent than the terminal ones.

Results and discussion

The influence of the mismatch position onto the duplex stability was evaluated. For this purpose five microarrays specific to five organisms were prepared to carry perfect matching and single-base mismatched probes. In fact, the formation of RNA-DNA duplexes is an equilibrium process during the hybridization. After the hybridization, the microarrays were very shortly washed at 20^{o}C with the solution of the same salt concentration as it was during the hybridization (5xSSC which is 1M sodium ions), thus preserving the state that had been achieved in the equilibrium. Apparently, a mismatch shifts the equilibrium to the extent depending on its position. Quantification shows (Figure 4-1) that our hypothetical function in average is followed, although one can not claim here that it is followed without exceptions.

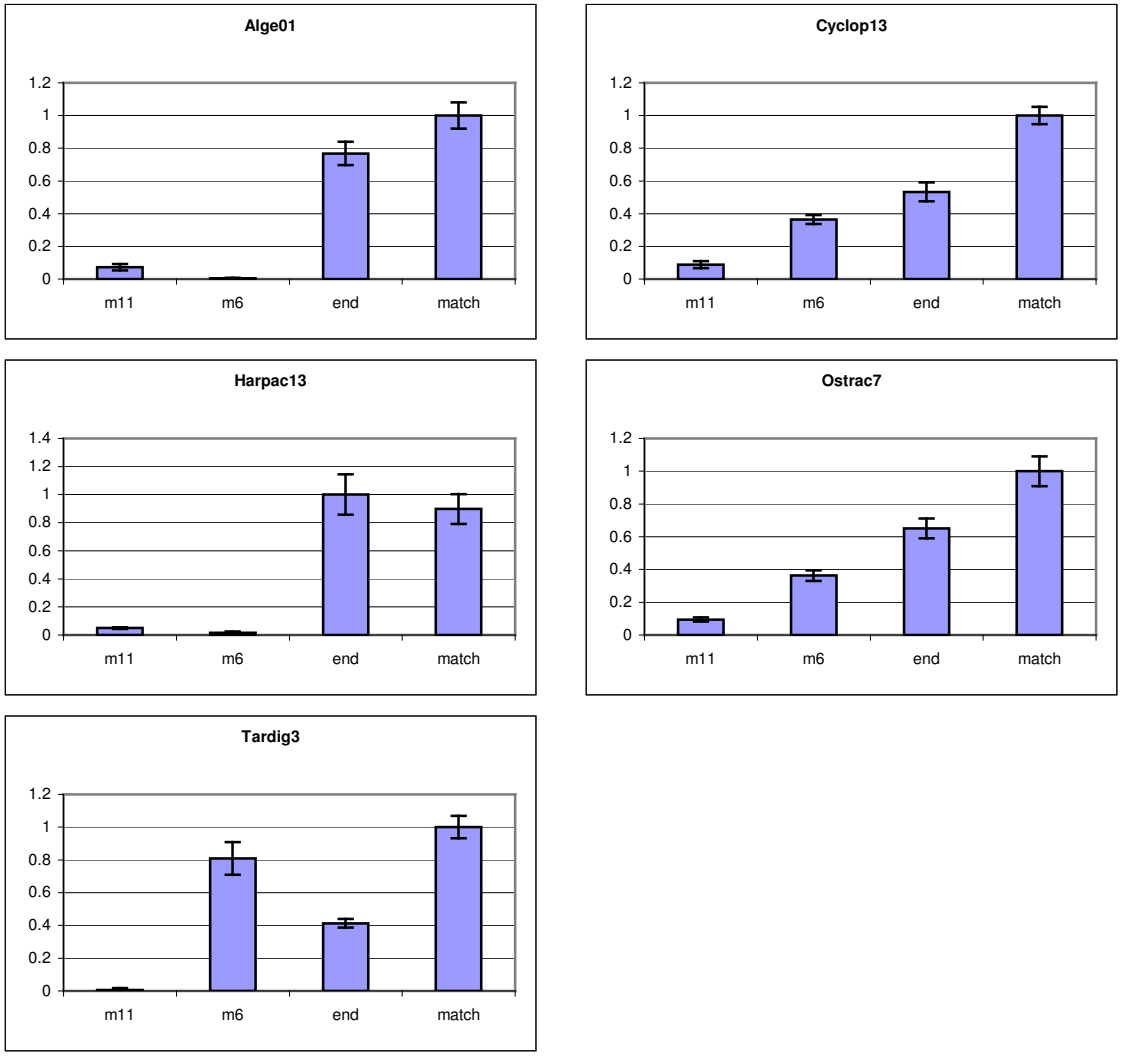


Figure 4-1. Influence of the mismatch position on the DNA-RNA duplex formation. Low stringency wash with 5xSSC, 20°C. Captions of each chart represent species applied onto the chips.

In all cases the most intermediate mismatch at the position eleven brings the highest instability of the duplex. It is necessary to point out that based on the nearest neighbor model the standard way of stabilities prediction fails to calculate the experimentally revealed behavior (Figure 4-2).

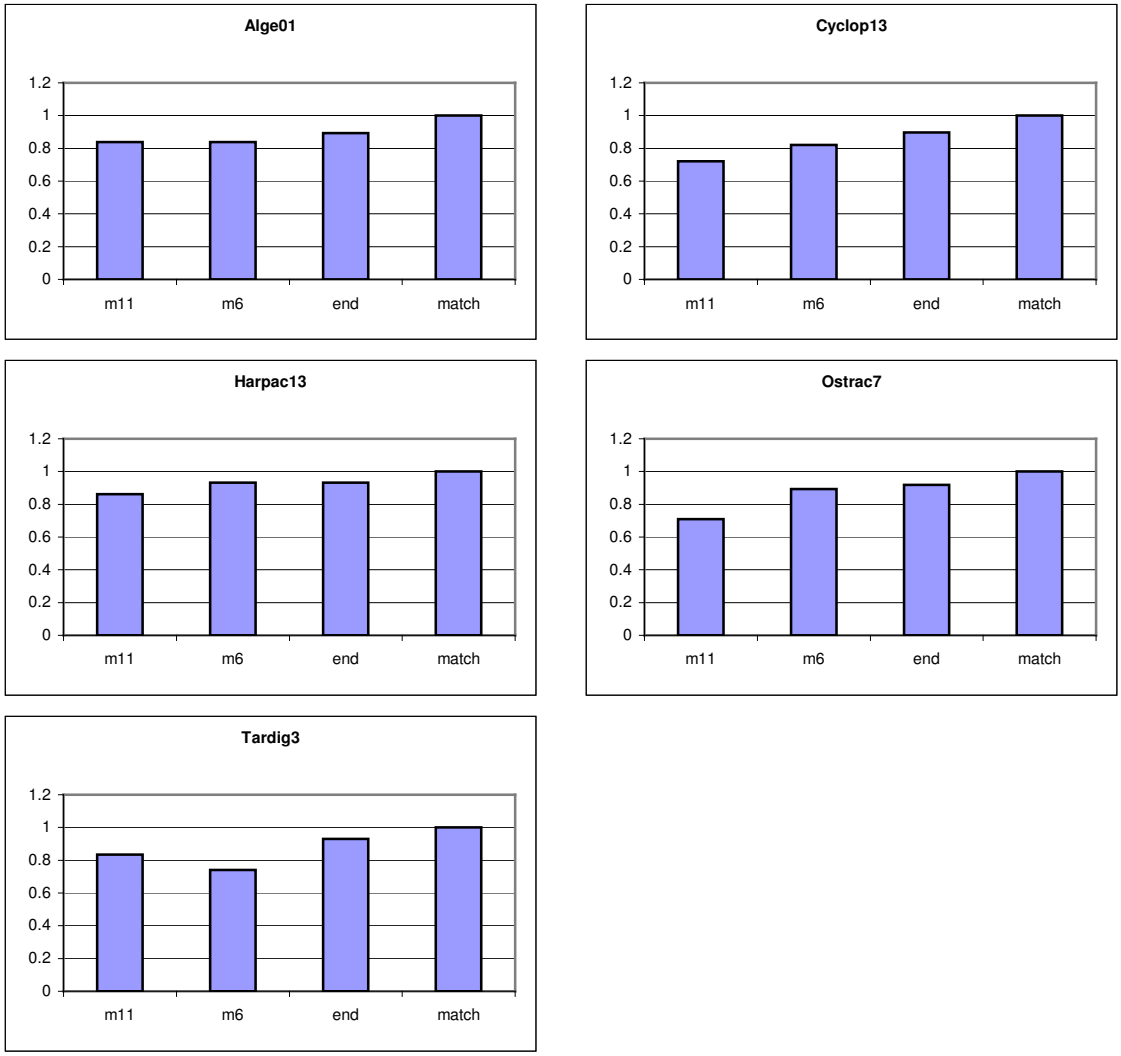


Figure 4-2. Theoretically calculated concentrations of duplexes for the species indicated as captions of the charts.

The standard way predicts for the studied probes melting temperatures more than 55 ^{o}C slightly varying from probe to probe. The Materials and Methods section provides means how to determine concentrations of the duplexes predicted by the nearest neighbor model at any experimental temperature, and these normalized concentrations are shown in Figure 4-2. The reason for such a discrepancy between the actual situation and prediction given by the nearest neighbor model might be attributed to the surface effects that to some extent depend on the density of the immobilized probes [81,82]. Moreover, there is a study indicating that the nearest neighbor parameters found for solution conditions are not directly applicable to the chips [83].

The 5xSSC washing is naturally unlikely to be used in practice; I tried a more traditional wash with 0.1xSSC. Quantification shows much more contrasted dependency (Figure 4-3). The decreased ionic strength enhances repulsion of the DNA backbone and in general destabilizes DNA duplex and already mismatched DNA duplexes are disturbed even further. These results demonstrate the practical applicability of the dissociation model proposed in our work [64] and moreover, the stability predictions are robust towards the salt concentration.

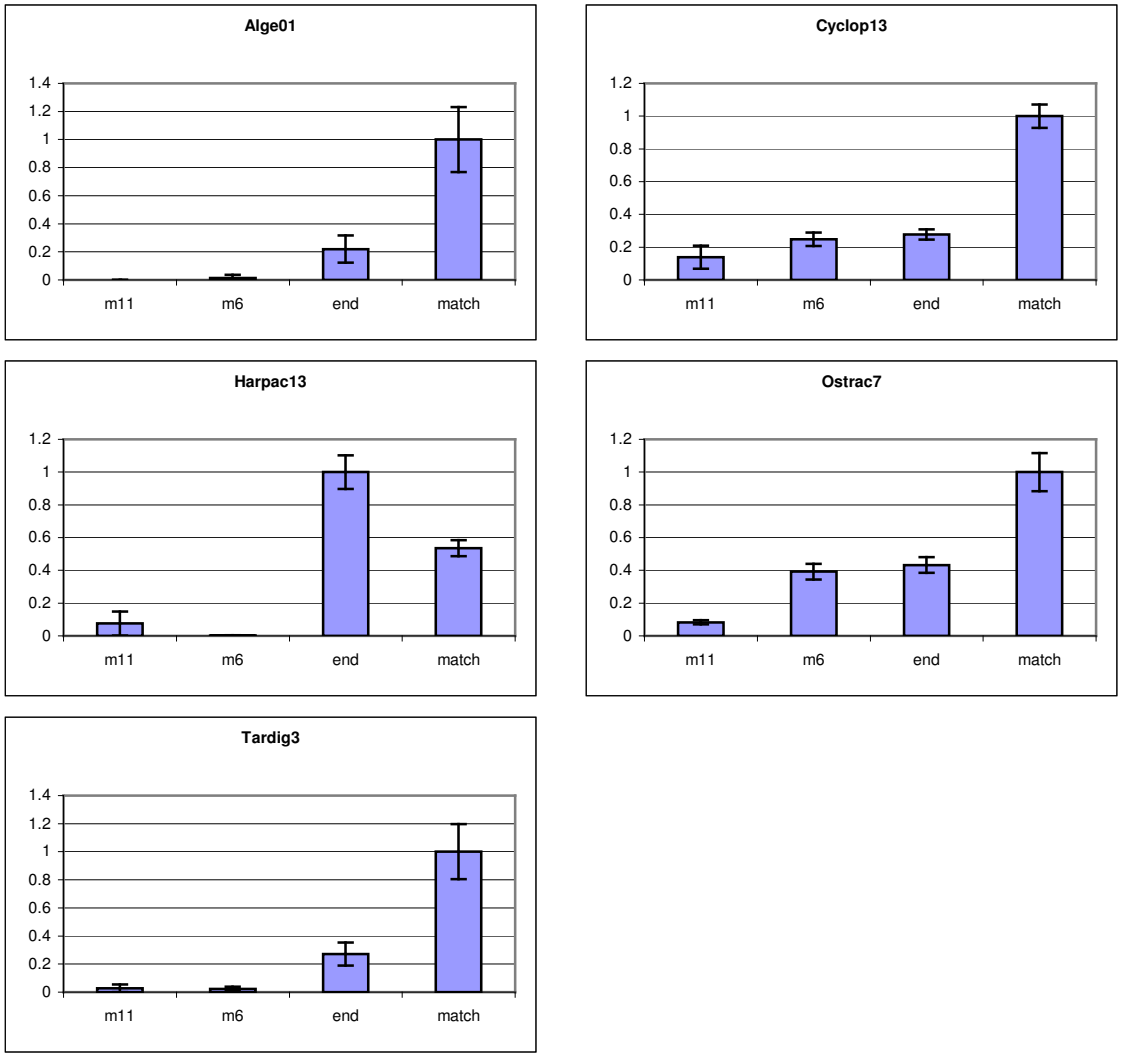


Figure 4-3. Influence of the mismatch position on the DNA-RNA duplex formation. High stringency wash with 0.1xSSC, 20^{o}C. Captions of each chart represent species applied onto the chips.

Another issue addressed was how accurate the PROBE proposes organism specific probes. The array containing ten probes for ten different organisms, five of which being a negative control, was studied with five rRNAs separately. Recognition power of such an array is high enough, as one can see on Figure 4-4. Experimental conditions were the same as for the mismatch position assay.

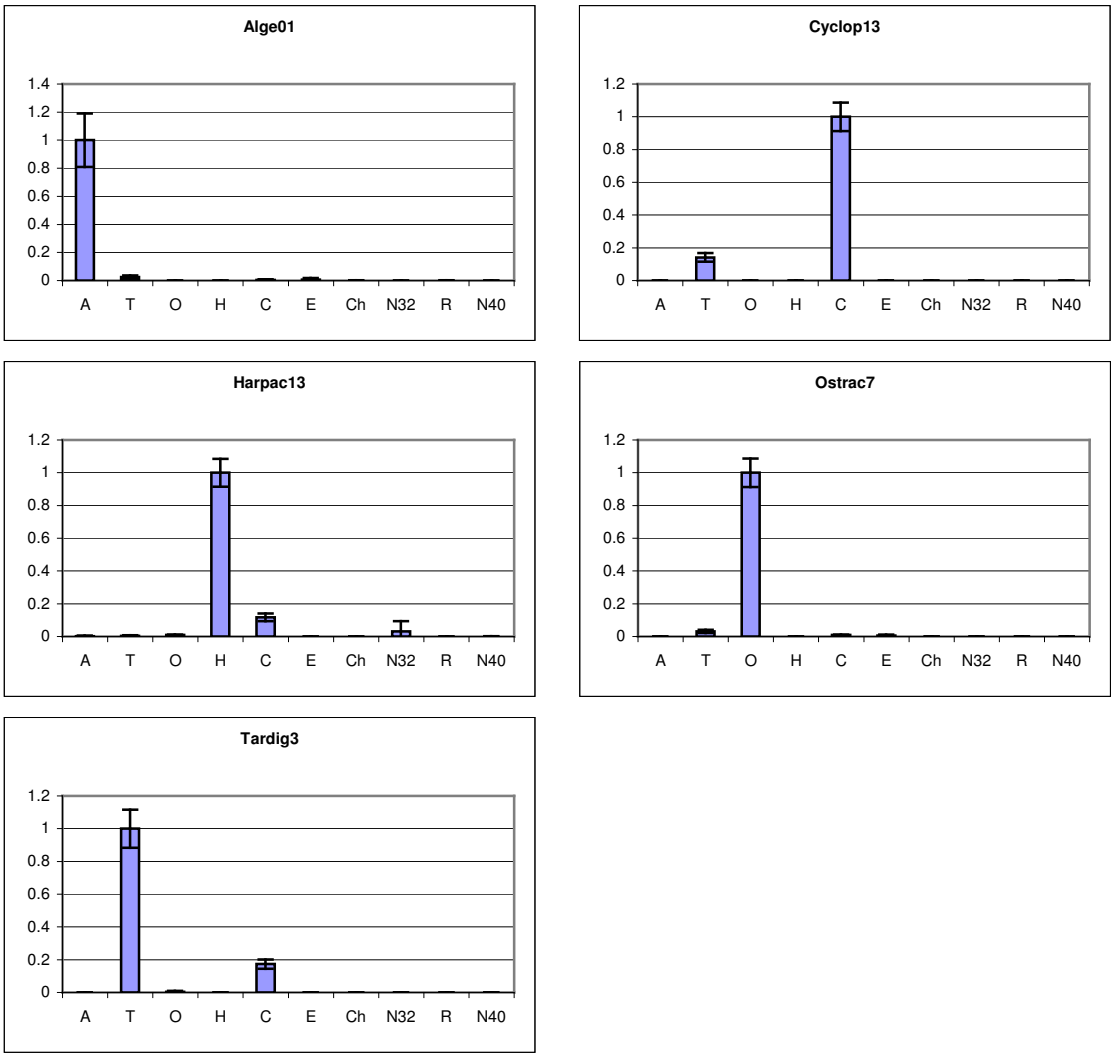


Figure 4-4. Recognition of RNA species by microarray. Horizontal axis represents spots available on the microarray specific to the following species: A – Alge01, T – Tardig3, O – Ostrac7, H – Harpac13, C – Cyclop13, E – Epheme1, Ch – ChiRon14, N32 – Nematd32, R – Rotato06, N40 – Nematd40. Low stringency wash with 5xSSC, 20°C. Captions of each chart represent species applied onto the chips.

One can see that the highest signal appears at the spots specific to a given rRNA species being applied onto the microarrays. Along with the specific signal there is a certain level of cross-hybridization. To increase the specificity even further, the microarrays were washed with 0.1xSSC. There is an obvious gain in the specificity as one can see on Figure 4-5.

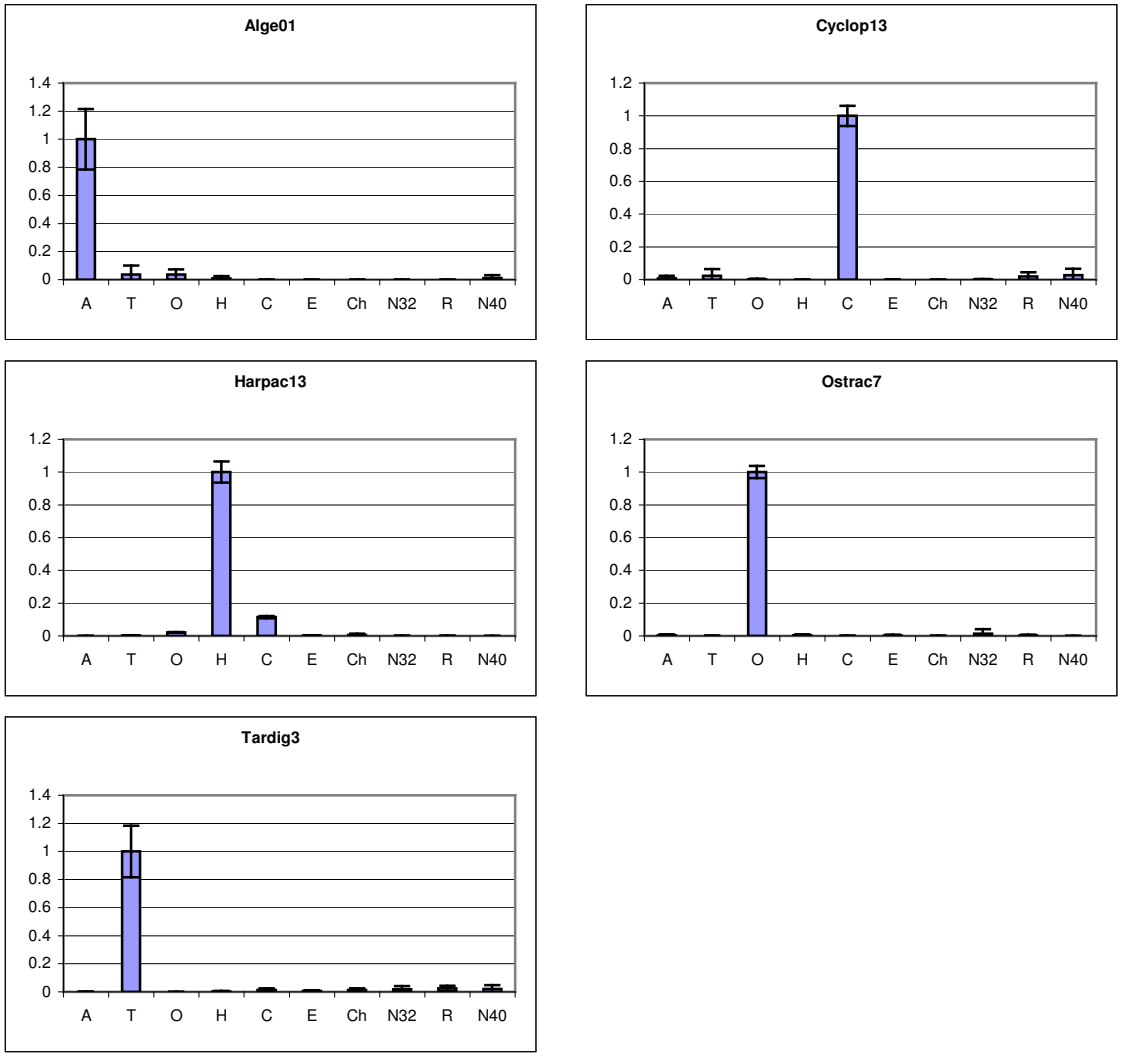


Figure 4-5. Recognition of RNA species by microarray. Horizontal axis represents spots available on the microarray specific to the following species: A – Alge01, T – Tardig3, O – Ostrac7, H – Harpac13, C – Cyclop13, E – Epheme1, Ch – ChiRon14, N32 – Nematd32, R – Rotato06, N40 – Nematd40. High stringency wash with 0.1xSSC, 20°C. Captions of each chart represent species applied onto the chips.

A correct quantification is only possible when the response of the microarray on the various target concentrations is known. I studied the response of the microarrays on the increasing concentration of RNA. In this experiment two RNA species (Harpac13 and Osrtrac7) were in the mixtures – one as a standard (Harpac13), another one was varied (Ostrac7). Altogether 6 hybridizations on separate microarrays were performed with varying Ostrac7 RNA concentration of 0.1, 1, 10, 100 and 500 times standard. One can see on Figure 4-6 that the microarray responds hyperbolically.

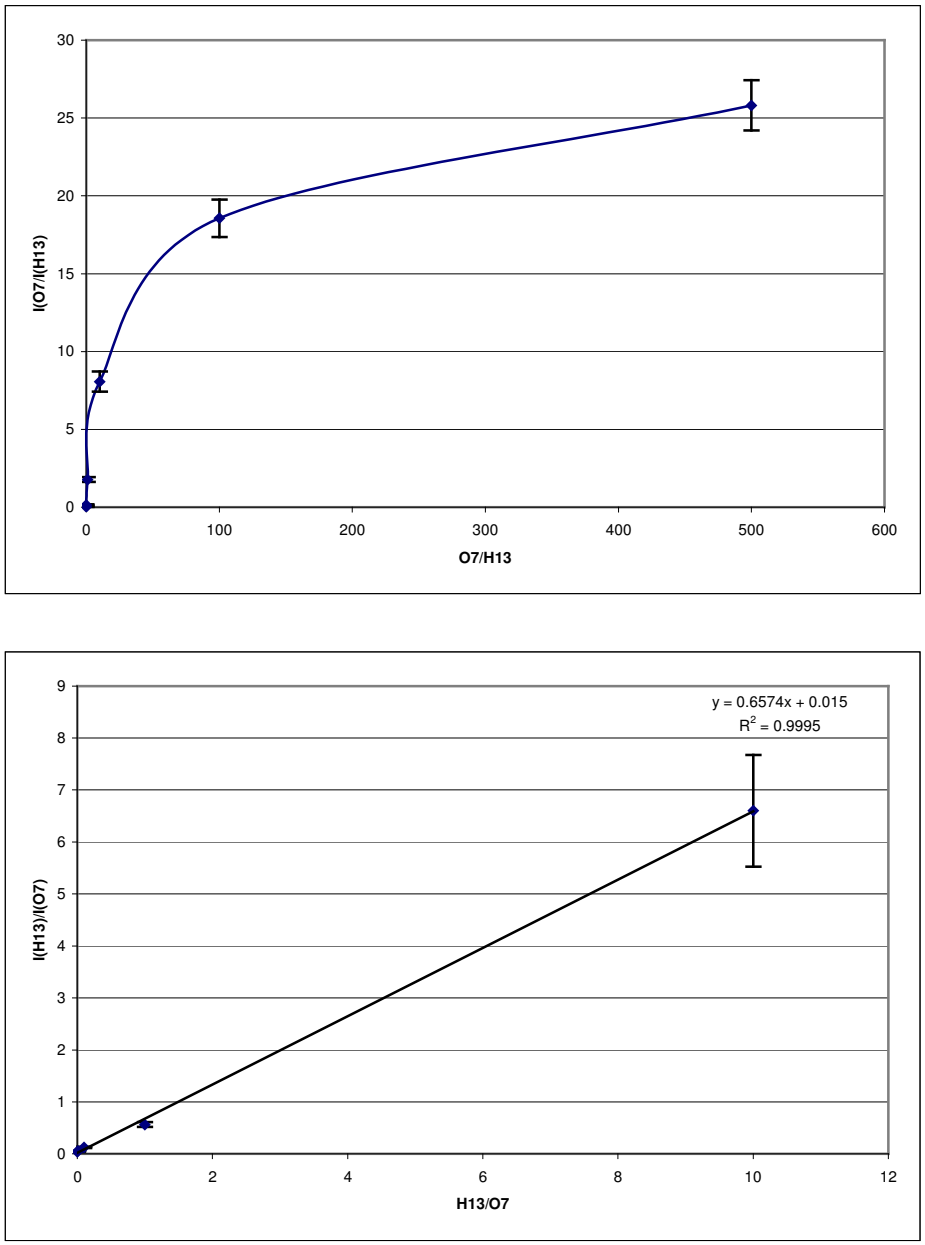


Figure 4-6. Microarray response calibration. Lengmuir's model.

This is a classical example of Langmuir's absorption model, which assumes an absorptive surface with a limited number of absorption centers. The microarray is indeed a sort of a surface with a limited number of hybridization (say absorption) centers. To test if it is true, I have performed coordinates transformation (power -1) and tested if the obtained curve is a line. In fact, one can see that it is indeed a line and the goodness of fit is very high (0.999). The same behavior was observed for GeneChips (Affymetrix) where the hybridization at high concentrations followed the Langmuir's curve [85]. In the mentioned above experiments (mismatched probes and evaluation of the probes specificity), the range of concentrations was far away from

the saturation. The concentrations correspond to 1 at Figure 4-6. Therefore, the response of the microarray can be considered linear and levels of fluorescence are directly proportional to the duplex amounts, i. e. stabilities.

Conclusion

The influence of position of mismatches on the duplex stability has been experimentally demonstrated in the equilibrium conditions that are assumed by the model [64]. Experimentally it has been shown that the PROBE designs probes of desired specificity without any significant cross hybridization.

Materials and methods

Computation methods

The oligos employed in the microarray experiments were designed using the PROBE [64] being given with a database of Benthos organisms and other eucaryotes [80]. The search was performed assuming the dissociation probability of 24% within the group being detected, and 76% outside this group. The oligonucleotides of approx. 50% GC composition were selected from the pool of designed probes.

For comparison with the nearest neighbor model, the prediction of melting temperatures according to that model were performed by TermAlign software [67]. Predicted melting temperature determines the overall stability of the duplex, the calculation of the predicted quantities of duplexes at experimental conditions (Figure 4-2) was performed as follows.

Let the duplex formation be designated by the equilibrium:

$$\mathrm{A} + \mathbf{A} \leftrightarrows \mathrm{A}\mathbf{A}$$

where A is one strand, **A** is the complementary one. This equilibrium is characterized by the equilibrium constant $Kp = [\mathrm{A}\mathbf{A}]/([\mathrm{A}][\mathbf{A}])$; T, p = const. Thus, a well-known relation holds [84]:

$$\Delta G^0 = -RT\ln(Kp)$$

Equation 4-1. Standard Gibbs energy change at isobaric isothermic conditions.

At the melting point [A**A**] = [A] = [**A**] = $C_0/2$, where C_0 is the total concentration of the double stranded polynucleotide. Thus, the equilibrium constant at the melting point is equal to $1/(C_0/2)$. This enables to calculate the predicted ΔG^0 for each duplex according to Equation 4-1. To calculate the predicted equilibrium constant appearing at the experimental conditions Kp^{ex} with temperature less than the melting one, the following relation is used:

$$-RT_{melt}\ln\left(\frac{1}{C_0/2}\right) = -RT_{ex}\ln\left(Kp^{ex}\right),$$

therefore

$$\ln\left(Kp^{ex}\right) = \frac{T_{melt}}{T_{ex}}\ln\left(\frac{1}{C_0/2}\right).$$

To compare the actual quantities of duplexes a system of three equations was solved comprised from the relations for definition of the equilibrium constant, material balance and equality of strands concentrations. The solution is as follows:

$$[\mathbb{A}\mathbf{A}] = \left(\frac{1}{\sqrt{Kp^{ex}}} - \sqrt{\frac{1+Kp^{ex}C_0}{Kp^{ex}}}\right)^2$$

Experimental procedures

The microarrays were created using MWG Epoxy slides (MWG). The 5' C6-amino modified oligonucleotides (Metabion) were delivered onto the slides by the Gene TAC spotting robot (Genomic Solutions). Immobilization of 5' oligonucleotides was performed at 42 ^oC at 50% humidity for 8-12 hours. The necessary humidity was established in a 1L glass of 180 mm height containing 100 ml of water. The microarray slides were put at 40 mm elevation above the water level. The glass was tightly sealed with several sheets of filter paper. By variation of the number of the filter paper sheets it was possible to set up any desired humidity. In fact, the filter paper determines a vapor flow at a certain rate, which in turn determines a gradient of humidity along the altitude of the glass. By variation of the rate, the steepness of the gradient can be varied, and thus the humidity at a certain elevation.

The plasmids of the pZErO-2 vector carrying rRNA genes were used to transform the TOP10 *E coli* strain by electroporation according to a standard protocol [79]. The transformed cells were grown on the LB agar plates containing kanamycin. LB liquid medium (30 ml) containing 40 μg/ml kanamycin was inoculated with single colonies of each clone. After overnight growth at 37 ^{o}C, the cells were transferred to 470 ml of LB medium containing 40 μg/ml kanamycin the cells were grown overnight at 37 ^{o}C. The harvest was used to isolate the plasmids according to the large scale plasmid preparation and alkaline lysis protocols [79].

Large amounts of ribosomal RNA were produced by *in vitro* transcription reaction using the RiboMAX (Promega) kit containing either Sp6 or T7 polymerases depending on the orientation of inserted rRNA genes. Again, depending on the orientation of the insert, the plasmids before transcription were first cut with either SpeI or XbaI (Roche) restriction enzymes according to the cognate protocol. To determine the orientation of the inserts, the plasmids were sequenced with M13 Forward and Reverse primers. The sequencing was performed on the MegaBACE sequencer (Molecular Dynamics) according to its standard protocol.

Hybridization experiments were carried out as follows. Reaction mixture contained 10 μl of approx. 17 nM (5 ng/μl) rRNA under investigation, 100 μl of hybridization solution (5xSSC, 0.2 mg/ml BSA, 0.1%SDS), 6 μl of 200 mM Ribonucleoside Vanadyl Complex (New England Biolabs) and 5 μl of 100 μM "marking oligo" (see the next section) `5′-Cy3-CTC-CTT-GGT-CCG-TGT-TTC-AAG-ACG-G-3′`. Thus, the final concentration of RNA was 1.4 nM. The reaction mixture was applied onto the chip heated up to 70 ^{o}C in the Gene TAC Hybridization Station (Genomic Solutions), then the system was heated up to 80 ^{o}C, incubated for 1 minute, then cooled down to 45 ^{o}C and incubated with agitation for 24 hours. The chips were washed with 5xSSC or 0.1xSSC (depending on the experiment) at 20 ^{o}C 3 times.

Scanning was performed on the Gene TAC LS IV scanner (Genomic Solutions) with an appropriate laser and a filter set for Cy3. After the first scanning the chips were scanned again with a high gain. These scans showed "black holes" on a bright

background at the positions of the spots where hybridization did not happen (see Figure 4-7).

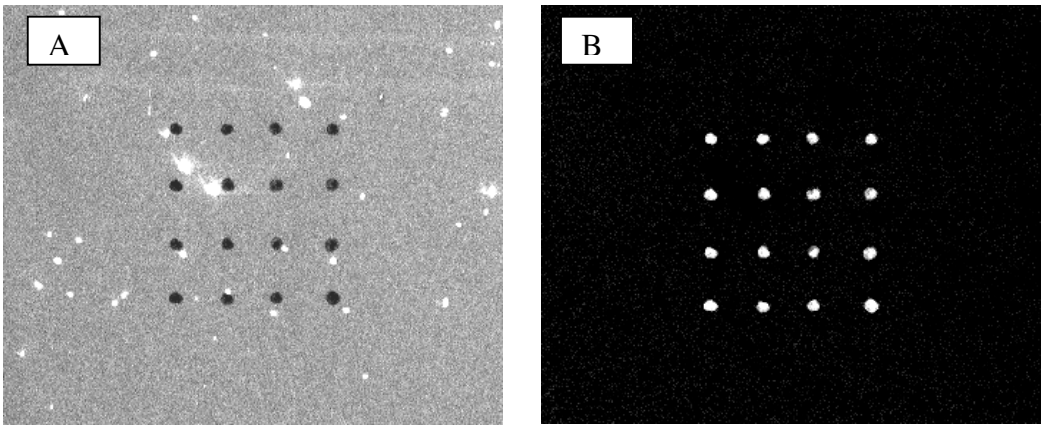


Figure 4-7. Display of spots without hybridization. A) **- "black holes".** B) **- negated and contrasted "black holes".**

This phenomenon has been used to find out the positions of all spots no matter how strong the hybridization signal is. The images were negated, contrasted and processed with software that automatically determines spot positions. After the spot positions were found it became possible to quantify signals of any intensity. All images were analyzed with ArrayVision 7.0 (Imaging Research).

Indirect labeling

In order to simplify labeling of RNA and at the same time increase the accuracy of quantification by having a single label per single RNA molecule, an indirect labeling method was used. A fluorescently labeled oligonucleotide – the “marking oligo” – hybridizes with a conserved region of RNA, while all the variable regions are available for hybridization with specific oligonucleotides of a microarray. The determined detection limit of the method is not worse than 300 pM of rRNA (60 pg/μl).

Tables

Table 4-1. Perfectly matching probes

Organism	Sequence
Tardig3	CAC–CAC–CAA–GTT–ACG–GGA–TC
Nematd40	TGC–ATT–CGC–GAG–AGT–GTC–TG
Rotato6	TAC–TAA–GCT–CTC–TGC–CGA–CG
Nematd32	CCA–ACA–GAG–TTG–ACC–TTT–AC
ChiRon14	GGA–TGG–AAG–TGG–CGA–CTG–TT
Epheme1	ACA–TTC–GAC–CGA–CTA–GCG–GA
Cyclop13	ACA–TAC–ATG–GAT–CAC–CCC–TC
Harpac13	CTT–TGA–CAG–CGA–TCA–CCC–CT
Ostrac7	ATC–ACT–CGC–GCA–TAA–GTT–AG
Alge1	GAT–TGT–GCA–ATA–CTC–CAA–CC

Table 4-2. Mismatched probes

	Alge1
end	GAT–TGT–GCA–ATA–CTC–CAA–C**T**
m6	GAT–TGT–GCA–ATA–CT**A**–CAA–CC
m11	GAT–TGT–GCA–**C**TA–CTC–CAA–CC
	Cyclop13
end	ACA–TAC–ATG–GAT–CAC–CCC–T**T**
m6	ACA–TAC–ATG–GAT–CA**T**–CCC–TC
m11	ACA–TAC–ATG–**A**AT–CAC–CCC–TC
	Harpac13
end	CTT–TGA–CAG–CGA–TCA–CCC–C**C**
m6	CTT–TGA–CAG–CGA–TC**G**–CCC–CT
m11	CTT–TGA–CAG–**T**GA–TCA–CCC–CT
	Ostrac7
end	ATC–ACT–CGC–GCA–TAA–GTT–A**A**
m6	ATC–ACT–CGC–GCA–TA**G**–GTT–AG
m11	ATC–ACT–CGC–**A**CA–TAA–GTT–AG
	Tardig3
end	CAC–CAC–CAA–GTT–ACG–GGA–T**T**
m6	CAC–CAC–CAA–GTT–AC**A**–GGA–TC
m11	CAC–CAC–CAA–**A**TT–ACG–GGA–TC

Chapter 5 Quantification of a Mixed Sample by Sequencing

Introduction

Environmental samples are always complex mixtures of different organisms. For the taxonomical studies one has to quantify each individual organism. One of the quantification ways is to determine the concentration of characteristic nucleic acids of the organism, for example rRNA. Traditional means of discovering a mixture composition rely on the separation of the mixture to the individual components and their subsequent quantification or using specific probes recognizing each component in the presence of the others. Here we propose another method based on the sequencing of the complex mixture provided that the sequences of the characteristic nucleic acids of the suspected organisms are known.

Solution

Sequencing can be performed by various means, but the most appropriate for our problem is pyrosequencing. Pyrosequencing is a new technique based on the fact that the DNA polymerization reaction is followed by a pyrophosphate release. During sequencing one adds a certain nucleotide tri-phosphate to the mixture of a template, polymerase and an annealed primer. If the incorporation occurs, the pyrophosphate is released. The pyrophosphate along with luciferin and luciferase produce a spark of light that can be measured and quantified. The amount of light is proportional to the concentration of the pyrophosphate and thus to the concentration of the template.

The problem of finding out the amounts of each individual sequence species can be represented as a system of linear equations:

$$\begin{cases} \sum_{i=1}^{N} k_{1i}(A) n_{1i}(A) x_i = S_1(A) \\ \sum_{i=1}^{N} k_{1i}(T) n_{1i}(T) x_i = S_1(T) \\ \sum_{i=1}^{N} k_{1i}(G) n_{1i}(G) x_i = S_1(G) \\ \sum_{i=1}^{N} k_{1i}(C) n_{1i}(C) x_i = S_1(C) \\ \dots \\ \sum_{i=1}^{N} k_{Li}(A) n_{Li}(A) x_i = S_L(A) \\ \sum_{i=1}^{N} k_{Li}(T) n_{Li}(T) x_i = S_L(T) \\ \sum_{i=1}^{N} k_{Li}(G) n_{Li}(G) x_i = S_L(G) \\ \sum_{i=1}^{N} k_{Li}(C) n_{Li}(C) x_i = S_L(C) \end{cases}$$

where S_j – peak intensity at j-th step of a specified nucleotide, $k_{ji}(X)$ – linear coefficient between brightness and incorporation event of a specified nucleotide X at the j-th step of sequencing for the i-th organism, $n_{ji}(X)$ – number of available incorporation events for the X nucleotide at the j-th step for the i-th organism (0,1,2,3…), x_i – the sought concentration of the desired i-th organism, L – number of steps.

In short this can be written in a matrix form:

$$\mathbf{N} \cdot \mathbf{X} = \mathbf{S}$$

where **N** – matrix of n_{ji} multiplied by $k_{ji}(X)$, **X** vector of x_i, **S** – vector of peak intensities. This system can be analytically solved [86].

First, multiplying both sides on $\mathbf{N^T}$:

$$\mathbf{N^T} \cdot \mathbf{N} \cdot \mathbf{X} = \mathbf{N^T} \cdot \mathbf{S}.$$

It is known that $\mathbf{N^T} \cdot \mathbf{N}$ is a square matrix no matter what **N** is, thus multiplying on the inverted $\mathbf{N^T} \cdot \mathbf{N}$ will produce a unity matrix:

$$\left(\mathbf{N^T} \cdot \mathbf{N}\right)^{-1} \cdot \mathbf{N^T} \cdot \mathbf{N} \cdot \mathbf{X} = \left(\mathbf{N^T} \cdot \mathbf{N}\right)^{-1} \cdot \mathbf{N^T} \cdot \mathbf{S}.$$

The matrix $(\mathbf{N^T \cdot N})^{-1} \cdot \mathbf{N^T \cdot N}$ is a unity matrix having 1 at its diagonal and 0 everywhere else. Hence, the solution is:

$$\mathbf{X} = \left(\mathbf{N}^{\mathbf{T}} \cdot \mathbf{N}\right)^{-1} \cdot \mathbf{N}^{\mathbf{T}} \cdot \mathbf{S}$$

and the diagonal of $(\mathbf{N^T \cdot N})^{-1}$ contains squares of standard deviations of each solution.

The number of steps required for an unambiguous solution must be at least as many as it provides non-singularity of the matrix **N**. In fact, it is better to overdefine the system of equations even further to make sure that glitches of the measured intensities do not affect the solution and the noise is averaged.

In practice the coefficients $k_{ji}(X)$ are unknown. In the ideal case they should be all equal to each other and thus may be omitted from the system. But the reality is different. To overcome the problem of unknown coefficients one has to record the pyrograms for each sequence – for example from the clones or synthesized oligonucleotides – and store them in a library. Therefore, a pyrogram of a given sequence is a column in the matrix **N.** The solution **X** is then found in folds of the concentrations used to record the library of pyrograms. Thus, it makes sense to use equal concentrations for recording of the library.

The library can be recorded once and stored for all further solutions. When sequencing the unknown mixture, it is good to add a known amount of a sequence that is not present in the sample, as a concentration standard. Naturally, the pyrogram of the standard must be in the library as well. After finding the solution, all the variables can be related to the standard and their final concentrations thus can be obtained via known concentration of the standard. Usage of the standard avoids sensitivity of the solution to the fluctuations of the instrument characteristics from one run to another.

The ideal case, where all $k_{ji}(X)$ being equal to 1, is necessary for determining the minimal number of steps for the sequencing depending on the actual sequences. Apparently, the sequences per-se possess a certain amount of information and this influences the number of steps to perform. Therefore the matrix **N** contains only the numbers of nucleotides available for incorporation at each step. This matrix for a given number of steps can be generated from the sequences according to a simple algorithm. After the matrix has been generated the singularity must be tested and if it

is singular, more steps are added, matrix re-generated, tested again and so on until it is not singular anymore.

Experimental Verification

The experimental verification of the proposed method was carried out with PCR products of cloned rRNA genes. A library of pyrograms for 7 species has been recorded. The abbreviations of the species are the following: A – Alge01, T – Tardig3, O – Ostrac7, H – Harpac13, C – Cyclop13, E – Epheme1, N4 – Nematd40. Figure 5-1 shows the profiles of the pyrograms.

The PCR products were discarded to simulate a real life experiment where the library is supposed to be pre-collected. Another PCR reaction with the same species has been carried out. Agarose gel electrophoresis reveals the concentration of the products to be equal. The estimated concentration was 40 ng/µl. The PCR products were used to create three mixtures. The Table 5-1 shows the composition of the mixtures in µl of the PCR products.

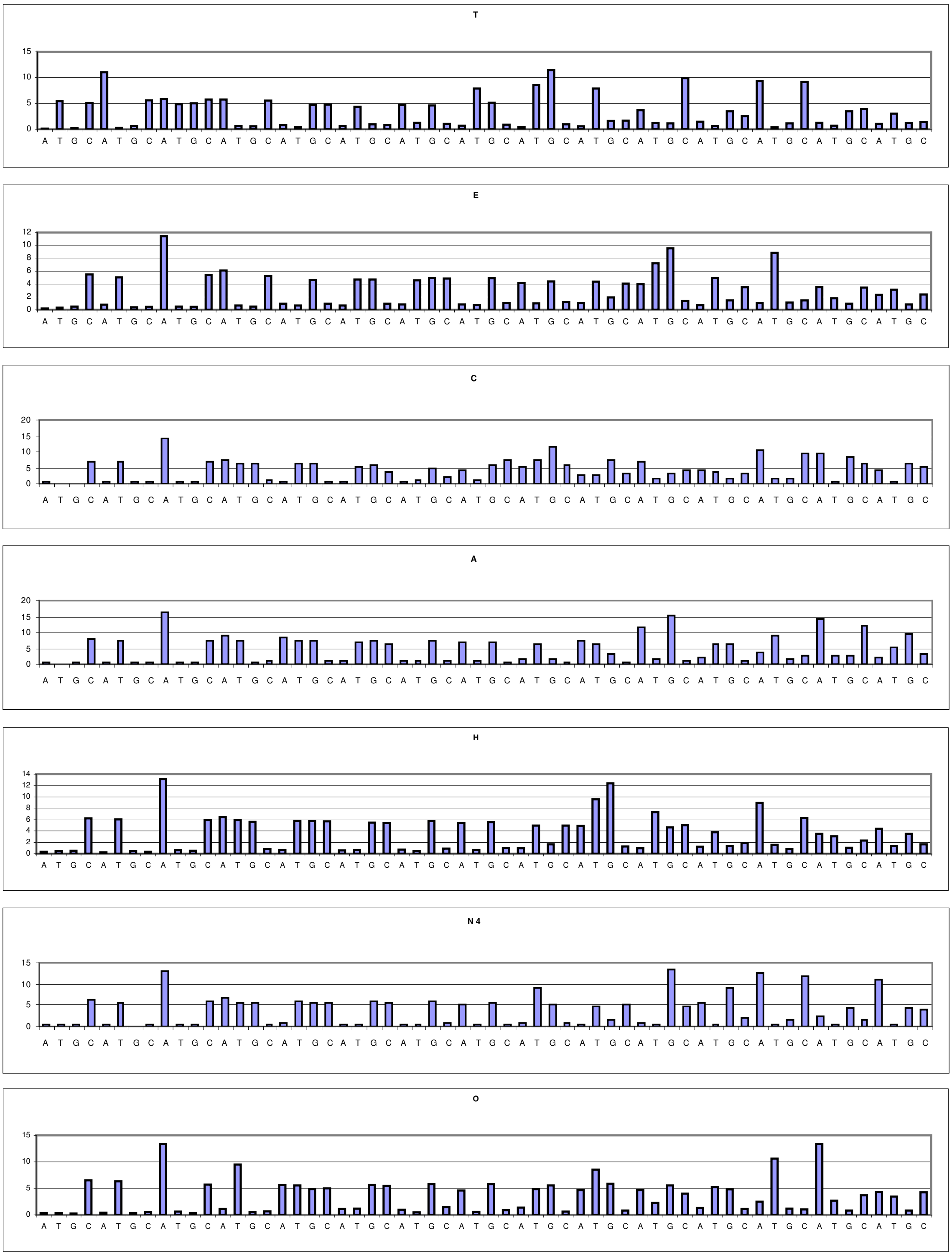


Figure 5-1. Library of pyrograms for species T,E,C,A,H,N4,O.

The mixtures were subjected to pyrosequencing and the obtained pyrograms are shown on Figure 5-2.

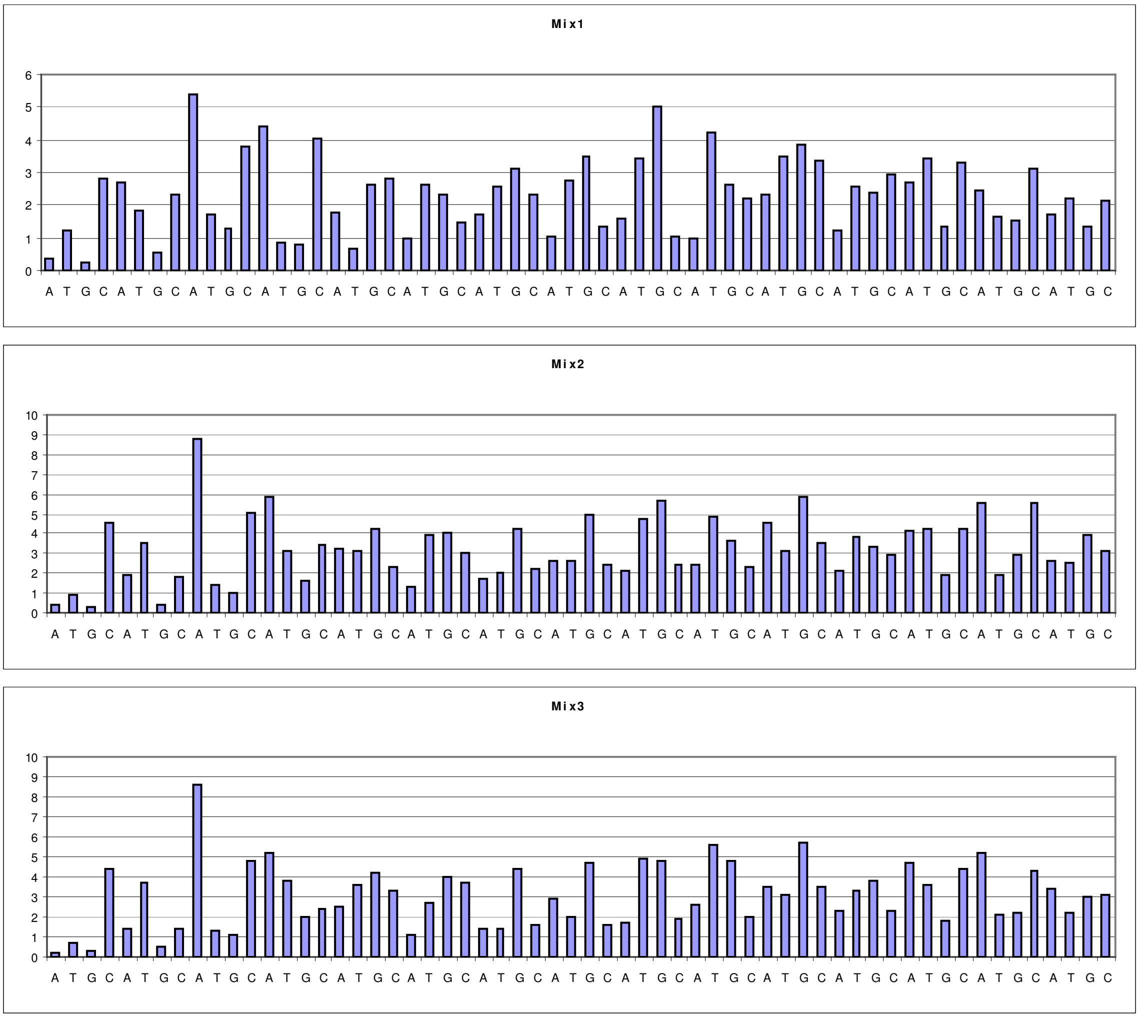


Figure 5-2. Pyrograms of the mixtures

These pyrograms were analyzed according to the mentioned above formalism. First, the analysis of the sequence information reveals the minimum number of steps to be 34. The pyrograms of the mixtures as well as the profiles from the library were truncated after the 50^{th} step thus providing 16 more equations to overdefine the system of equations. The system was solved and found variables were related towards "known" concentration of the standard, here the specie "T". The solutions and their standard deviations are shown in Table 5-2. Table 5-3 shows the final values. In Mix1 there is a good recognition of species. The absent species show values from –1 to 2, which is attributed to the background. Indeed, the standard deviations for these solutions were larger than the solutions itself, while that for the solutions of present species were naturally less than the discovered values. Analogously, the appropriate recognition of the compositions of the other mixtures Mix2 and Mix3 is possible (see Table 5-3).

Conclusion

Sequencing of a complex mixture of DNA species theoretically makes it possible to discover quantities of each component provided that the nucleotide sequence of each species is known. This has been proven practically by deciphering of several mixtures of various composition. The approach is proposed for quantification of mixtures of small organisms from environmental samples.

Materials and Methods

The plasmids of the pZErO-2 vector carrying rRNA genes were prepared as described in the Materials and Methods section of the Chapter 4.

PCR amplification was carried out in 30 µl volume with 37 cycles. Concentration of the template 0.67 ng/µl. The primer 5′-GAC-CCG-TCT-TGA-AAC-ACG-G-3′ was used as a forward primer, the 5'biotynilated primer 5′-ATC-GAT-TTG-CAC-GTC-AGA-A-3′ was used as a reverse primer.

Pyrosequencing was performed by the PSQ 96 instrument (Pyrosequencing AB, Sweden) with 5′-GAA-ACA-CGG-ACC-AAG-GAG-T-3′ sequencing primer. Prior to sequencing the PCR products were cleaned up according to the protocol provided by the pyrosequencer manual.

The solution of linear equations was carried out with the Matlab package.

Tables

Table 5-1. Three mixtures

Specie	Mix1	Mix2	Mix3
T	10	10	10
E	13	10	7
C		7	3
A		9	5
H			6
N4			4
O			5

Table 5-2. Solutions and standard deviations

	Mix1	Mix2	Mix3	St. Dev.
T	0.2717	0.1991	0.1535	0.0436
E	0.3513	0.2378	0.1346	0.0663
C	-0.0098	0.1498	0.0705	0.0616
A	-0.0226	0.1588	0.0997	0.0678
H	0.0522	0.0469	0.1609	0.0685
N4	-0.0026	-0.0031	0.0759	0.0608
O	0.0517	0.0493	0.1174	0.0837

Table 5-3. Comparison between found and given ratios of the species. T is a concentration standard.

	Mix1		Mix2		Mix3	
	given	found	given	found	given	found
T	10	10	10	10	10	10
E	13	13	10	12	7	9
C	0	-0	7	7	3	4
A	0	-1	9	8	5	6
H	0	2	0	2	6	10
N4	0	-0	0	-0	4	5
O	0	2	0	2	5	8

References

1. DJ. Lockhart, H. Dong, MC. Byrne, MT. Folletti, MV. Gallo, MS. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, EL. Brown. **Expression monitoring by hybridization to high-density oligonucleotide arrays.** *Nat. Biotechnol.*, 1996, **14:** 1675–1680.

2. M. Schena, D. Shalon, R. Heller, A. Chai, PO. Brown and RW. Davis. **Parallel human genome analysis: microarray-based expression monitoring of 1000 genes.** *Proc. Natl. Acad. Sci. USA*, 1996, **93:** 10614–10619.

3. AC. Pease, D. Solas, EJ. Sullivan, MT. Cronin, CP. Holmes, SPA. Fodor. **Light-generated oligonucleotide arrays for rapid DNA sequence analysis.** *Proc. Natl. Acad. Sci. USA*, 1994, **91:** 5022–5026.

4. J. Khan, LH. Saal, ML. Bittner, Y. Chen, JM. Trent, PS. Meltzer. **Expression profiling in cancer using cDNA microarrays.** *Electrophoresis*, 1999, **20:** 223–229.

5. S. Drmanac, D. Kita, I. Labat, B. Hauser, C. Schmidt, JD. Burczak and R. Drmanac. **Accurate sequencing by hybridization for DNA diagnostics and individual genomics.** *Nat. Biotechnol.*, 1998, **16:** 54–58.

6. G. Yershov, V. Barsky, A. Belgovskiy, E. Kirillov, E. Kreindlin, I. Ivanov, S. Parinov, D. Guschin, A. Drobishev, S. Dubiley, A. Mirzabekov. **DNA analysis and diagnostics on oligonucleotide microchips.** *Proc. Natl. Acad. Sci. USA*, 1996, **93:** 4913–4918.

7. JG. Hacia, FS. Collins. **Mutational analysis using oligonucleotide microarrays.** *J. Med. Genet.*, 1999, **36:** 730–736.

8. DR. Call, DP. Chandler, FJ. Brockman. **Fabrication of DNA microarrays using unmodified oligomer probes.** *BioTechniques*, 2001, **30:** 368–379.

9. A. Spiro, M. Lowe, D. Brown. **A bead-based method for multiplexed identification and quantitation of DNA sequences using flow cytometry.** *Appl. Environ. Microbiol.*, 2000, **66:** 4258–4265.

10. D. Guschin, B. Mobarry, D. Proudnikov, D. Stahl, Brittman, A. Mirzabekov. **Oligonucleotide microchips as genosensors for determinative and environmental studies in microbiology.** *Appl Environ Microbiol* 1997, **63:** 2397-2402.

11. R. Amann, W. Ludwig. **Ribosomal RNA-targeted nucleic acid probes for studies in microbial ecology.** *FEMS Microbiol Rev* 2000, **24**: 555-565.

12. R. Amann, W. Ludwig, K. Schleifer. **Phylogenetic identification and in situ detection of individual microbial cells without cultivation.** *Microbiol Rev* 1995, **59**: 143-169.

13. E. DeLong, G. Wickham, N. Pace. **Phylogenetic stains: ribosomal RNA-based probes for the identification of single cells.** *Science* 1989, **243**: 1360-1363.

14. The ARB project. http://www.arb-home.de/ .

15. H. Allawi, A and J SantaLucia Jr.. **Thermodynamics and NMR of Internal G•T Mismatches in DNA.** *Biochemistry* 1997, **36**: 10581-10594.

16. N. Peyret, P. Senevirante, H. Allawi, J. SantaLucia Jr.. **Nearest-Neighbor Thermodynamics and NMR of DNA Sequences with Internal A•A, C•C, G•G and T•T Mismatches.** *Biochemistry* 1999, **38**: 3468-3477.

17. S. Fodor, R. Liphutz, X. Huang. **The Robert A Welch Foundation 37th Conference on Chemical Research 40 Years of the DNA Double Helix.** Houston, Texas, October 25-26, 1993.

18. M. Leijon, A. Gräslund. **Effects of sequence and length on imino proton exchange and base pair opening kinetics in DNA oligonucleotide duplexes.** *Nucleic Acids Res* 1992, **20**: 5339-5343.

19. DJ. Patel, SA. Kozlowski, S. Ikuta, K. Itakura. **Deoxyadenosine-deoxycytidine pairing in the d(C-G-C-G-A-A-T-T-C-A-C-G) duplex: conformation and dynamics at and adjacent to the dA X dC mismatch site.** *Biochemistry* 1984, **23:** 3218-26.

20. S. Bavykin, V. Mikhaylovich, V. Zakharyev, Yu. Lysov, J. Kelly, J. Flax, J. Jackman, D. Stahl, A. Mirzabekov. **Discrimination of Bacillus anthractis and closely related organisms by analysis of 16S and 23S rRNA with oligonucleotide microchips.** *Appl Environ Microbiol* 2001, in press.

21. M. Kalnik, D. Norman, B. Li, P. Swann, D. Patel. **Conformational transitions in thymidine bulge-containing deoxytridecanucleotide duplexes. Role of flanking sequence and temperature in modulating the equilibrium between looped out and stacked thymidine bulge states.** *J Biol Chem* 1990, **265**: 636-647.

22. BL. Maidak, JR. Cole, TG. Lilburn, CT. Parker Jr, PR Saxman, RJ. Farris, GM. Garrity, GJ. Olsen, TM. Schmidt, JM. Tiedje. **The RDP-II (Ribosomal Database Project).** *Nucleic Acids Res* 2001, **29**: 173-174.

23. W. Melchior Jr, P. von Hippel. **Alteration of the relative stability of dA-dT and dG-dC base pairs in DNA.** *Proc Natl Acad Sci USA* 1973, **70**: 298-302.

24. M. Doktycz, M. Morris, S. Dormady, K. Beattie, B. Jacobson. **Optical melting of 128 octamer DNA duplexes. Effects of base pair location and nearest neighbors on thermal stability.** *Journal of Biological Chemistry* 1995, **270**: 8439-8445.

25. C. Petzold. **Programming Windows. 5th edition**. 1998, Microsoft Press.

26. Nancy Winnick Cluts. **Programming the Windows 95 User Interface (Microsoft Programming).** 1995, Microsoft Press.

27. **Platform SDK.** MSDN Library, Visual C++ 6.0, Microsoft.

28. B. Stroustroup. **The C++ Programming Language (3rd Edition).** 1997, Addison-Wesley Pub Co.

29. K. Brockschmidt. **Inside OLE (Microsoft Programming) 2nd Book.** 1995, Microsoft Press.

30. H. Dai, M. Meyer, S. Stepaniants, M. Ziman, R. Stoughton. **Use of hybridization kinetics for differentiating specific from non-specific binding to oligonucleotide microarrays.** *Nucleic Acids Res.*, 2002, **30**: e86.

31. AL. Drobyshev, C. Machka, M. Horsch, M. Seltmann, V. Liebscher , MH. de Angelis, J. Beckers. **Specificity assessment from fractionation experiments (SAFE): a novel method to evaluate microarray probe specificity based on hybridisation stringencies.** *Nucleic Acids Res.*, 2003, **31:** e1.

32. S. Ikuta, K. Takagi, R.B. Wallace, K. Itakura. **Dissociation kinetics of 19 base paired oligonucleotide-DNA duplexes containing different single mismatched base pairs.** *Nucleic Acids Res*., 1987, **15**: 797-811.

33. S. Wang, AE. Friedman, ET. Kool. **Origins of high sequence selectivity: a stopped-flow kinetics study of DNA/RNA hybridization by duplex- and triplex-forming oligos.** *Biochemistry*, 1995, **34:** 9774–9784.

34. N. Tibanyenda, SH. De Bruin, CA. Haasnoot, GA. van der Marel, JH. van Boom, CW. Hilbers, **The effect of single base-pair mismatches on the duplex stability of d(T-A-T-T-A-A-T-A-T-C-A-A-G-T-T-G).d(C-A-A-C-T-T-G-A-T-A-T-T_A-A-T-A).** *Eur. J. Biochem.*, 1984, **139:** 19–27.

35. AW-C. Liew, H. Yan, M. Yang. **Robust adaptive spot segmentation of DNA microarray images.** *Pattern Recognition*, 2003, **36:** 1251 –1254.

36. Y. Yang, H. Yan. **An adaptive logical method for binarization of degraded document images.** *Pattern Recognition*, 2000, **33:** 787-807.

37. P. Hegde, R. Qi, K. Abernathy, C. Gay, S. Dharap, R. Gaspard, J. Earle-Hughes, E. Snesrud, N. Lee, J. Quackenbush. **A Concise Guide to cDNA Microarray Analysis.** *BioTechniques*, 2000, **29**: 548-562.

38. R. Wünschiers, P. Lindblad. **Hydrogen in Education - A Biological Approach.** *Int. J. Hydrogen Energy*, 2002, **27**: 1131-1140.

39. J. Small, DR. Call, FJ. Brockman, TM. Straub, DP. Chandler. **Direct Detection of 16S rRNA in Soil Extracts by Using Oligonucleotide Microarrays.** *Applied and Environmental Microbiology*. 2001, **67**: 4708-4716.

40. JL. DeRisi, VR. Iyer PO. Brown.. **Exploring the metabolic and genetic control of gene expression on a genomic scale.** *Science*, 1997, **278**: 680–686.

41. M. Schena, D. Shalon, RW. Davis, PO. Brown. **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science*, 1995, **270**: 467–470.

42. K. Wang, L. Gan, E. Jeffery, M. Gayle, AM. Gown, M. Skelly, PS. Nelson, WV. Ng, M. Schummer, L. Hood, J. Mulligan. **Monitoring gene expression profile changes in ovarian carcinomas using cDNA microarray.** *Gene*, 1999, **229**: 101–108.

43. M. Wilson, J. DeRisi, H. Kristensen, P. Imboden, S. Rane, PO. Brown, GK. Schoolnik. **Exploring drug-induced alterations in gene expression in Mycobacterium tuberculosis by microarray hybridization.** *Proc. Natl. Acad. Sci. USA*, 1999, **96**: 12833–12838.

44. GP. Yang, DR. Ross, WW. Kuang, PO. Brown, RJ. Weigel. **Combining SSH and cDNA microarrays for rapid identification of differentially expressed genes.** *Nucleic Acids Res.*, 1999, **27**: 1517–1523.

45. KL. Gunderson, XC. Huang, MS. Morris, RJ. Lipshutz, DJ. Lockhart, M. Chee. **Mutation detection by ligation to complete n-mer DNA arrays.** *Genome Res.*, 1998, **8**: 1142–1153.

46. L. Wodicka, H. Dong, M. Mittmann, M. Ho, DJ. Lockhart. **Genome-wide expression monitoring in *Saccharomyces cerevisiae*.** *Nat. Biotechnol.*, 1997, **15**: 1359–1367.

47. D. Guiliano, M. Ganatra, J. Ware, J. Parrot, J. Daub, L. Moran, H. Brennecke, JM. Foster, T. Supali, M. Blaxter, AL. Scott, SA. Williams, BE. Slatko.

Chemiluminescent detection of sequential DNA hybridizations to high-density, filter-arrayed cDNA libraries: a subtraction method for novel gene discovery. *BioTechniques*, 1999, **27**: 146–152.

48. GT. Hermanson. **Bioconjugate techniques.** 1996, Academic Press, Inc., San Diego, Calif.

49. M. Chee, R. Yang, E. Hubbell, A. Berno, XC. Huang, D. Stern, J. Winkler, DJ. Lockhart, MS. Morris, SPA. Fodor. **Accessing genetic information with high-density DNA arrays.** *Science*, 1996, **274**: 610–614.

50. MT. Cronin, RV. Fucini, SM. Kim, RS. Masino, RM. Wespi, CG. Miyada. **Cystic fibrosis mutation detection by hybridization to light-generated DNA probe arrays.** *Hum. Mutat.*, 1996, **7**: 244–255.

51. PN. Gilles, DJ. Wu, CB. Foster, PJ. Dillon, SJ. Chanock. **Single nucleotide polymorphic discrimination by an electronic dot blot assay on semiconductor microchips.** *Nat. Biotechnol.*, 1999, **17**: 365–370.

52. TR. Gingeras, G. Ghandour, E. Wang, A. Berno, P M. Small, F. Drobniewski, D. Alland, E. Desmond, M. Holodniy, J. Drenkow. **Simultaneous genotyping and species identification using hybridization pattern recognition analysis of generic Mycobacterium DNA arrays.** *Genome Res.*, 1998, **8**: 435–448.

53. JG. Hacia, CB. Lawrence, MS. Chee, SPA. Fodor, FS. Collins. **Detection of heterozygous mutations in BRCAI using high density oligonucleotide arrays and two-color fluorescence analysis.** *Nat. Genet.*, 1996, **14**: 441–447.

54. MJ. Kozal, N. Shah, N. Shen, R. Yang, R. Fucini, TC. Merigan, DD. Richman, D. Morris, E. Hubbell, M. Chee, TR. Gingeras. **Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays.** *Nat. Med.*, 1996, **2**: 753–759.

55. DJ. Lockhart, H. Dong, MC. Byrne, MT. Follettie, MV. Gallo, MS. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, EL. Brown. **Expression monitoring by hybridization to high-density oligonucleotide arrays.** *Nat. Biotechnol.*, 1996, **14**: 1675–1680.

56. RJ. Sapolsky, RJ. Lipshutz. **Mapping genomic library clones using oligonucleotide arrays.** *Genomics*, 1996, **33**: 445–456.

57. S. Tyagi, D. P. Bratu, FR. Kramer. **Multicolor molecular beacons for allele discrimination.** *Nat. Biotechnol.*, 1998, **16**: 49–53.

58. A. de Saizieu, U. Certa, J. Warrington, C. Gray, W. Keck, J. Mous. **Bacterial transcript imaging by hybridization of total RNA to oligonucleotide arrays.** *Nat. Biotechnol.*, 1998, **16**: 45–48.

59. D. Proudnikov, A. Mirzabekov. **Chemical methods of DNA and RNA fluorescent labeling.** *Nucleic Acids Res.*, 1996, **24**: 4535–4532.

60. SG. Bavykin, JP. Akowski, VM. Zakhariev, VE. Barsky, AN. Perov, AD. Mirzabekov. **Portable System for Microbial Sample Preparation and Oligonucleotide Microarray Analysis.** *Applied and Environmental Microbiology*, 2001, **67**: 922–928.

61. Molecular Probes Handbook. http://www.probes.com/handbook/.

62. E. Kowalski. **Nuclear Electronics.** 1970, Springer Verlag, Berlin, Heidelberg, New York.

63. Tiff images processing library. http://www.libtiff.org.

64. A Pozhitkov, D Tautz: **An algorithm and program for finding sequence specific oligo-nucleotide probes for species identification.** *BMC Bioinformatics* 2002 3:9.

65. F. Li, G. Stromo. **Selection of Optimal DNA Oligos for Gene Expression Arrays.** *Bioinformatics*, 2001, **17**: 1067-76.

66. KE. Ashelford, AJ. Weightman, JC. Fry. **PRIMROSE: a Computer Program for Generating and Estimating the Phylogenetic Range of 16S rRNA Oligonucleotide Probes and Primers in Conjunction with the RDP-II Database.** *Nucleic Acids Res.*, 2002, **30**: 3481-3489.

67. L. Kaderali, A Schliep. **Selecting Signature Oligonucleotides to Identify Organisms Using DNA Arrays.** *Bioinformatics*, 2002, **18**: 1340-1349.

68. J-M. Rouillard, M. Zuker, E. Gulari. **OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach.** *Nucleic Acids Research*, 2003, 31: 3057-3062.

69. OV. Matveeva, SA. Shabalina, VA. Nemtsov, AD. Tsodikov, RF. Gesteland, JF. Atkins. **Thermodynamic calculations and statistical correlations for oligo-probes design.** *Nucleic Acids Research*, 2003, **31**: 4211-4217.

70. J. SantaLucia Jr., HT. Allawi, PA. Seneviratne. **Improved Nearest-Neighbor Parameters for Predicting DNA Duplex Stability.** *Biochemistry*, 1996, **35**: 3555-3562.

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

72. HT. Allawi, J. SantaLucia Jr. **Nearest Neighbor Thermodynamic Parameters for Internal G·A Mismatches in DNA.** *Biochemistry*, 1998, **37**: 2170-2179.

73. HT. Allawi, J. SantaLucia Jr. **Nearest-Neighbor Thermodynamics of Internal A·C Mismatches in DNA: Sequence Dependence and pH Effects.** *Biochemistry*, 1998, **37**: 9435-9444.

74. HT. Allawi, J. SantaLucia Jr. **Thermodynamics of internal C·T mismatches in DNA.** *Nucleic Acids Research*, 1998, **26**: 2694–2701.

75. DA. Stahl, R. Amann. **Development and application of nucleic acid probes.** p. 205–248. In E. Stackebrandt and M. Goodfellow (ed.), Nucleic acid techniques in bacterial systematics. 1991, John Wiley & Sons Ltd., Chichester, United Kingdom.

76. H. Urakawa, PA. Noble, S. El Fantroussi, JJ. Kelly, DA. Stahl. **Single-Base-Pair Discrimination of Terminal Mismatches by Using Oligonucleotide**

Microarrays and Neural Network Analyses. *Applied and Environmental Microbiology*, 2002, **68**: 235–244.

77. H. Urakawa, S. El Fantroussi, H. Smidt, JC. Smoot, EH. Tribou, JJ. Kelly, PA. Noble, DA. Stahl. **Optimization of Single-Base-Pair Mismatch Discrimination in Oligonucleotide Microarrays.** *Applied and Environmental Microbiology*, 2003, **69:** 2848–2856.

78. B. Persson, K. Stenhag, P. Nilsson, A. Larsson, M. Uhlen, PA. Nygren. **Analysis of Oligonucleotide Probe Affinities Using Surface Plasmon Resonance: A Means for Mutational Scanning.** *Analytical Biochemistry*, 1997, **246:** 34–44.

79. T. Maniatis, J. Sambrook, EF. Fritsch. **Molecular Cloning: A Laboratory Manual.** 1989, Cold Spring Harbor Laboratory; 2nd Edition/3 Volume Set edition.

80. M. Markmann. PH. D theses, 1999, University of Munich.

81. AW. Peterson, LK. Wolf, RM. Georgiadis. **Hybridization of Mismatched or Partially Matched DNA at Surfaces.** *J. Am. Chem. Soc.*, 2002, **124**: 14601-14607.

82. A. Vainrub, BM. Pettitt. **Surface Electrostatic Effects in Oligonucleotide Microarrays: Control and Optimization of Binding Thermodynamics.** *Biopolymers*, 2003, **68**: 265 –270.

83. A. Kunitsyn, S. Kochetkova, V. Florentiev. **Partial Thermodynamic Parameters for Prediction Stability and Washing Behavior of DNA Duplexes Immobilized on Gel Matrix.** *J. Biomol. Struct. Dyn.* 1996, **14**: 239-244.

84. W. Greiner, L. Neise, H. Stöcker. **Thermodynamics and Statistical Mechanics.** 2000, Springer Verlag.

85. D. Hekstra, AR. Taussig, M. Magnasco, F. Naef. **Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays.** *Nucleic Acids Research*, 2003, **31**: 1962-1968.

86. WH. Press, BP. Flannery, SA. Teukolsky, WT. Vetterling. **Numerical Recipes in C: The Art of Scientific Computing.** 1993, Cambridge University Press; 2nd edition.

Erklärung

Ich versichere, daß ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; daß diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; daß sie - abgesehen von unter angegebenen Teilpublikationen - noch nicht veröffentlicht worden ist sowie, daß ich eine solche Veröffentlichung vor Abschluß des Promotionsverfahrens nicht vornehmen werde.

Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. Diethard Tautz am Lehrstuhl für Evolutionsgenetik betreut worden.

Teilpublikationen

A Pozhitkov, D Tautz: **An algorithm and program for finding sequence specific oligo-nucleotide probes for species identification.** *BMC Bioinformatics* 2002 3:9.

Köln, den 13.09.2003 Unterschrift

Lebenslauf

Name	Pozhitkov Alexander
Geburtsdatum/-ort	09.06.1976 in Moskau, USSR
Staatsangehörigkeit	Russische Föderation
Familienstand	verheiratet
Eltern	Eugeniy Pozhitkov und Alla Savitscheva

Schulbildung:

1983 – 1991	Besuch einer Schule in Moskau (Russische Föderation)
1991 – 1993	Besuch einer Schule mit dem Schwerpunkt Chemie in Moskau (Russische Föderation)

Studium:

WS 1993/1994	Immatrikulation für den Diplomstudiengang Chemie an der Moskauer Staats Universität
30 Juni 1998	Erfolgreicher Abschluß der Hauptdiplomsprüfung in Chemie (Hauptfach Chemie der natürlichen Substanzen)
1998 – 2000	Wissenschaftliche Mitarbeiter am Institut für Enzymologie der Moskauer Staats Universität
WS 2000/2001	Beginn der Promotion am Lehrstuhl für Evolutionsgenetik bei Prof. Tautz an der Universität zu Köln mit dem Thema „Molecular Taxonomy. Bioinformatics and Practical Evaluation“
WS 2003/2004	Voraussichtlicher Abschluß der Promotion

Köln, den 13.09.2003 Unterschrift