

Predicting Crosshybridization in Microarrays

Lorenz Steinbock, Dr. Harald Simmler

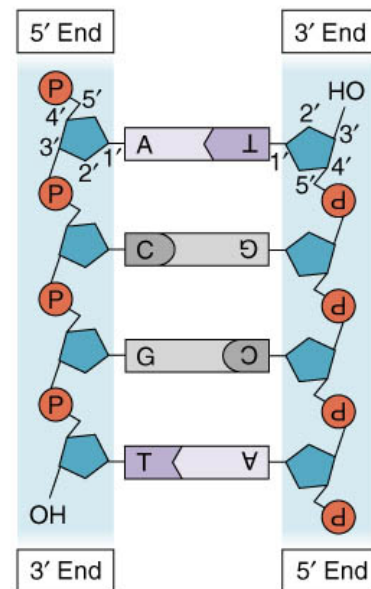
March 31, 2004, University of Heidelberg, Acconovis GmbH
eMail: lorenz@sibex.de, harald.simmler@acconovis.com

Abstract

Until today the algorithms for detection of segments in genes which can cause crosshybridization are limited, because they solve only one problem. In this work we will try to take into account the importance of similarity through BLASTn (Basic Local Alignment Search Tool nucleotide) and on the other hand, the thermodynamical effect approximated with the Nearest-Neighbour-Theory to forecast the chance of crosshybridization. By using these two tools we will show how probes which are likely to produce crosshybridization can be detected. Then we will show the proposed algorithm on yeast and one of its genes to illustrate the developed process.

1 Introduction

The analysis of the expression profile of a cell is important to understand which gene is up- or downregulated e.g. after incubation with a drug. Generally a gene is upregulated by copying the dsDNA (double stranded, ds) into ssmRNA (single stranded, ss) several times, whereas the mRNA or transcript is always synthesised from the 5' to 3' end (also see table 1). Then the mRNA is taken as a template to generate the protein. DNA-Microarrays have become a powerful tool to analyse the transcription profile of cells. The general principle is to immobilise many single ssDNA strands called "probes" of one single sequence on a distinct spot on a surface. These ssDNA strands are complementary to one gene which we want to detect. After repeating this procedure we obtain a microarray with different ssDNA strands on different spots (see figure 2). In the next step the cells which are of scientific interest are lysed and the mRNA is transformed using a reverse transcriptase into cDNA (single stranded). This cDNA is attached to a fluorescent protein and is called target. The entire cDNA is relished on the chip and after washing, only the cDNA which is complementary to the single strands on the chip should attach to the surface. To detect the cDNA, a



Copyright © 2001 Benjamin Cummings, an imprint of Addison Wesley Longman, Inc.

Figure 1: Scheme of DNA, [1]

laser is focused on the spot and the intensity of the emission is registered. The assumption that the intensity is proportional to the amount of cDNA attached to the spot permits us to calculate the number of bound cDNA.

Generally, we distinguish between oligonucleotide microarrays with only 20-60 nucleotides DNA strands and cDNA-Chips with DNA strands up to 300 nucleotides. The way these DNA-chips are made also differs from manufacturer to manufacturer. Affymetrix, the biggest company in this field, only sells oligonucleotide-microarrays with DNA strand lengths around 25 nucleotides [2]. Other firms sell the whole machine to fabricate the chips that are therefore called "custom-made" chips [3]. These DNA strands can vary from 20 to 60 nucleotides [4]. Unfortunately, the binding between the strand on the cDNA-Chip and the cDNA strand is not always selective which means that DNA strands and cDNA strands that are only 75% complementary to each other can also hybridise. This unspecific binding is also called crosshybri-

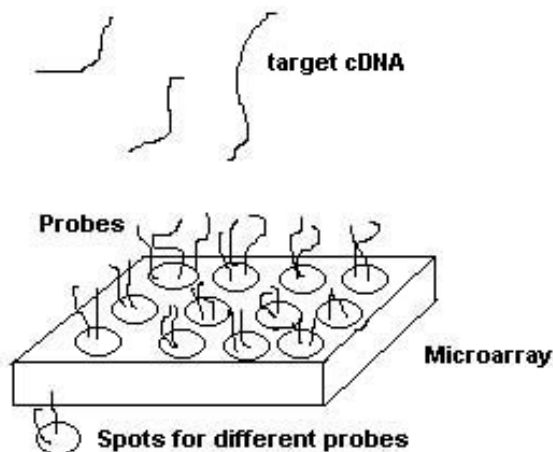


Figure 2: Scheme of a microarray

dization and the elimination of these errors is important to decrease the amount of false positive signals. The development of new algorithms for probe design is the most favoured way to eliminate false binding.

The presentation of an algorithm to discard probes which have a high chance to cause crosshybridization is the main goal of this paper.

2 State of the Art

There are currently two software types available, that are capable of both generating and increasing the quality of the signal.

2.1 Software Based on Sequence Databases

The first and most used type is based on sequence databases which require an entire genome sequence (see third and following programmes in Table 4). The goal is to detect unique regions in the gene, out of these regions we choose several probes, which are then checked against cross hybridization. First of all, the user has to choose which gene he wants to detect with the microarray and the organism in which the gene is expressed. Then the software compares the sequence of the gene with the whole genome and identifies unique parts in the gene which do not or do appear very rarely in the entire genome. This procedure is normally done by the subprogramme BLASTn, a heuristic algorithm which can "measure" the

amount of similarity (E-value) between different sequences. Another parameter to detect good probes is the minimum length of similarity, with a default value of 15 nucleotides in cDNA [5][6]. Further variables are the percentage of identity between possible probe and target which are allowed before the programme rejects the sequence. This parameter generally does not fall below 75% in cDNA [7]. Most of the literature refers to *Kane et al.* [8] for these default values, who examined these on 50mer oligonucleotide microarrays. Most programmes evaluate the position of the sequence and try to take probes from the 3' end to minimize the signal from half synthesised transcripts [9][10]. The melting temperature (T_m), the GC-percentage and the existence of low complexity regions are also used as parameters to evaluate different probes [11][12]. Furthermore, most of these programmes enable to "blast" the potentially unique sequences against the genome to detect probes that would cause crosshybridization.

2.2 Software Based on Thermodynamics

The second software type uses a more physical approach rather than a mathematical/ heuristic one to evaluate the possibility of crosshybridization (see first two programmes in Table 4). The programme "ChipCheck" from *Karsten et al.* [13] predicts the hybridization equilibria between probes and targets which can help to generate probes with low crosshybridization chances. To calculate the hybridization equilibria it uses the standard entropy, standard enthalpy and standard free enthalpy generated by HYTHER [14] which is based on the nearest-neighbour model developed by *Zimm et al.* [15]. Furthermore, it requires the strand concentration and the number of all probe molecules which are not always available. A similar solution is the commercial software "Sarani" [16] which estimates the melting temperature T_m using nearest-neighbour thermodynamics (explained later in the article) to calculate the minimum hybridization temperature to discard the non-specific cDNA strands. The advantage is that these approaches are based on physical properties. However, they generally require substantial computing power to calculate the hybridization equilibria for all targets and probes. Additionally, they do not take into account the

influence of crosshybridization with parts of the genome.

2.3 Bottom Line

In conclusion, we can say that all programmes of the first type use BLASTn or similar algorithms to quickly minimize the pool of possible probes and choose one of these strands depending on different parameters such as T_m or GC%. However, it should be mentioned that this way doesn't include the thermodynamical rules which determine how the strands hybridise. The second type of software takes physical measurements into account but requires substantial computing power to calculate the hybridization equilibria.

Therefore, a solution for these problems of missing physical foundation and limited computing resources should be a combination of these two approaches. In the first step, the gene is "blasted" against the genome to detect those regions in the gene which show the lowest similarity to other genes. Out of this minimized pool of strands we calculate with the nearest-neighbour-model the free energy between all probes and their crosshybridization partners. The probes with low free energies have a higher chance to cause crosshybridization than probes with high free energies and are not taken into account for the probe design.

3 Algorithm

3.1 BLASTn

BLASTn (Basic Local Alignment Search Tool nucleotide) is a fast heuristic method to search for regions of local similarity between sequences [17]. The main algorithm operates in three steps (see also Figure 3):

1. The sequence of the gene with n nucleotides is separated into parts of the length w . For example if $w = 8$, the first segment s_1 start with the first and terminate with the 8th nucleotide of the gene. The second segment s_2 would start with the second and end with the 9th nucleotide of the sequence. For a sequence a_h with $h = 100$ and $w = 20$, we get $n - w + 1 = 81$ different segments

$$a_1, a_2, \dots, a_{h-w+1}=81.$$

2. The $h - w + 1$ segments are called "words" and are now compared with the entire genome of the organism which contains the target gene. If we find the exact match of a w -mer in the database of the genome, we call it a word-hit.
3. These word-hits are extended dynamically in forward and reverse direction to . The extension is aborted if the score value S falls below a threshold T , if not, they are stored and called high scoring segment pairs (HSP).

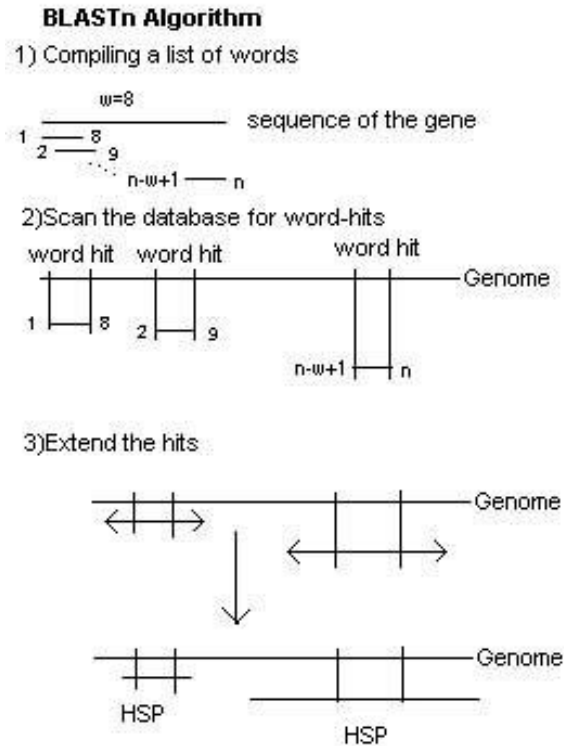


Figure 3: Summary of the BLASTn-Algorithm

The alignments are ranked by increasing E-values. The E-value is the number of times one expects to find alignments having score equal to or greater than S in a comparison of random sequences having the same length and composition as the query and the database:

$$E(\# \text{ of HSP's with } S > S_N) = \frac{N}{2^{S_N}}, \quad (1)$$

whereas S_N is a normalized score which depends on scoring parameters. $N = n \cdot m$ is the product of the genome and gene length.

Example

Equation (1) can be altered to

$$S_N = \log_2 \left(\frac{N}{E} \right)$$

which gives us the normalized score S_N to a specific E-value. If we had a gene with $n = 250$ nucleotides and a genome with $m = 50 \times 10^6$ nucleotides, we would expect for a E-value of 0.05 a normalized score S_N of 37.86. In conclusion, we can say that the lower the E-value is, the higher are the chances that the two sequences are similar.

3.2 NN-Thermodynamics

The Nearest-Neighbour Model is mainly pioneered by Tinoco and co-workers [19]. The function permits to calculate the free energy of a known sequence by considering (1) all nucleotides pairs and their following nucleotide pair, (2) the initial and ending nucleotide pair and (3) the possibility of selfcomplementary:

1. The contribution of each term for every pairwise interaction between base pairs. $\Delta G^\circ(i)$ are the standard free-energy changes for the 10 possible Watson-Crick NNs and n_i is the number of occurrences of each nearest neighbour: $\sum_i n_i \Delta G^\circ(i)$.
2. A positive free-energy ΔG penalty through the loss of translational freedom due to the formation of the first bonded base pair:
 $\Delta G^\circ(\text{init w/term } G \cdot C) + \Delta G^\circ(\text{init w/term } A \cdot T) = \Delta G_1^\circ$.
3. A positive free-energy penalty for self-complementary sequences ΔG_s° which is zero when the sequence is not self-complementary.

By uniting these three terms we get the following equation

$$\Delta G_t^\circ = \sum_i n_i \Delta G^\circ(i) + \Delta G_1^\circ + \Delta G_s^\circ$$

which represents the total free-energy of a definite duplex formation. By using the Gibbs-Helmholtz equation

$$\Delta G^\circ = \Delta H^\circ - T \Delta S^\circ$$

and the values of ΔH° and ΔS° (see table 1), which are found out by experiments, we can calculate ΔG_t° for every temperature.

| Sequence | $\Delta H^\circ \frac{\text{kcal}}{\text{mol}}$ | $\Delta S^\circ \frac{\text{cal}}{\text{mol K}}$ |
|-----------------|---|--|
| AA/TT | -7.9 | -22.2 |
| AT/TA | -7.2 | -20.4 |
| TA/AT | -7.2 | -21.3 |
| CA/GT | -8.5 | -22.7 |
| GT/CA | -8.4 | -22.4 |
| CT/GA | -7.8 | -21.0 |
| GA/CT | -8.2 | -22.2 |
| CG/GC | -10.6 | -27.2 |
| GC/CG | -9.8 | -24.4 |
| GG/CC | -8.0 | -19.9 |
| init w/term G·C | 0.1 | -2.8 |
| init w/term A·T | 2.3 | 4.1 |
| sym | 0.0 | -1.4 |

Table 1: Thermodynamic Parameters for DNA Helix Initiation and Propagation [14]

Example

$$\begin{aligned}
 & 5' \text{ T-A-G-C-T-A } 3' \\
 & 3' \text{ A-T-C-G-A-T } 5' \\
 T & = 50 C^\circ \equiv 323.15 K \\
 \Delta G_t^\circ & = \Delta G^\circ(\text{TA/AT}) + \Delta G^\circ(\text{AG/TC}) \\
 & \quad + \Delta G^\circ(\text{GC/CG}) + \Delta G^\circ(\text{CT/GA}) \\
 & \quad + \Delta G^\circ(\text{TA/AT}) + \Delta G^\circ(\text{init}) \\
 & \quad + \Delta G^\circ(\text{sym}) \\
 \Delta G_t^\circ & = [(-7200 - 323.15 \cdot -21.3) \\
 & \quad + (-7800 - 323.15 \cdot -21) \\
 & \quad + (-9800 - 323.15 \cdot -24.4) \\
 & \quad + (-7800 - 323.15 \cdot -21) \\
 & \quad + (-7200 - 323.15 \cdot -21.3) \\
 & \quad + (2300 - 323.15 \cdot 4.1) \\
 & \quad + (2300 - 323.15 \cdot 4.1) \\
 & \quad + (0.0 - 323.15 \cdot -1.4)] \text{ cal/mol} \\
 \Delta G_t^\circ & = [-316.905 - 1013.85 - 1915.14 \\
 & \quad - 1013.85 - 316.905 + 975.085 \\
 & \quad + 975.085 + 452.41] \text{ cal/mol} \\
 & = 2,17 \text{ kcal/mol}
 \end{aligned}$$

In conclusion, we can suppose that the duplex formations with lower ΔG_t° than other duplexes have a greater stability. The analogy would be a chemical reaction which has a higher chance to occur when the free energy is reduced by increasing the temperature [14].

3.3 Mathematical Model

The input sequence consists of one target sequence J_t which we want to detect and the whole genome G_m which can be approximated by one long sequence. The gene has t positions $1, 2, \dots, t$, whereas the genome has m positions. Each position p_1, \dots, p_t of J_t and p_1, \dots, p_m of G_m contains one nucleotide A, C, G or T. Comparing the sequence J_t with the Genome G_m by using BLASTn, we filter those segments out which show a low E-Value because they would indicate a strong similarity. Only the segments with an E-value above a certain threshold E_{thresh} are retained because they would probably form duplexes with the genome with the lowest stability:

$$S_u \in J_t : E(S_u) > E_{\text{thresh}}.$$

S_u represents u segments s_1, s_2, \dots, s_u which are all part of the target sequence J_t . The segments $S_u = s_1, \dots, s_u$, in combination with their counterparts in the sequence of the genome form a number of double layers.

$D_x = d_1, \dots, d_x$, whereas x equals the number u . Each double helix d_1, \dots, d_x represents a sequence of paired nucleotides with y positions p_1, \dots, p_y , whereas each position contains two nucleotides or even only one nucleotide:

$$\begin{aligned} &AA, AT, TT, CC, CG, GG, AG, AC, \\ &CT, GT, C, G, A, T. \end{aligned}$$

It is important to notice that besides the Watson-Crick-Pairing AT and CG, there are also false pairings like CC, GA or even a gap which are a consequence of choosing only sequence pairings with high E-values. These exceptions are not taken into account for the thermodynamical analysis and are considered as no binding because their contribution is very low or the thermodynamical data are not available.

Example

An example of a double helix d could be:

$$\begin{array}{cccccccc} \dots & p_3 & p_4 & p_5 & p_6 & p_7 & \dots & \\ 5' \dots & T & G & A & T & A & \dots 3' & \text{Part of } J_t \\ & | & | & & | & & & \\ 3' \dots & A & C & A & A & G & \dots 5' & \text{Part of } G_m. \end{array}$$

Each d_1, \dots, d_u is used as input sequence for the following thermodynamical examination:

$$\Delta G_t^{\circ}(d) = \sum_i n_i \Delta G^{\circ}(i) + \Delta G_1^{\circ} + \Delta G_3^{\circ}.$$

Out of this pool the duplex with the lowest free-energy $\Delta G_t^{\circ}(d_u)$ is defined as d_u^* :

$$\Delta G_t^{\circ}(d_u^*) = \max[\Delta G_t^{\circ}(D_u)].$$

d_u^* represents the duplex formation between the genome G_m and a segment out of the target sequence J_t which has in total the lowest free-energy and is therefore the most stable duplex. Hence, this probe would probably crosshybridize with the genome, it will not be used to detect the gene.

Pseudocode

```
for a Gene  $J_t$  :
do BLASTn ( $J_t$  against  $G_m$ ) =  $E(s_1), \dots$ 
  if  $E(s_i) > E_{\text{thresh}}$  store  $s_i$ 
  if  $\Delta G_t(d_i) < \text{min}\Delta G$ 
    set  $\text{min}\Delta G = \Delta G_t(d_i)$ , store  $d_i, s_i$ 
  end with  $E(s_u)$ 
store  $\text{min}\Delta G_t^{\circ}(d_u^*), d_u^*$ 
```

4 Test Implementation of the Model

To illustrate the proposed algorithm we will apply it to the following example: We are doing a gene expression analysis on the organism *Saccharomyces cerevisiae* (yeast) and we want to detect the gene SEO1 (Suppressor of Sulfoxide Ethionine resistance). Therefore we have to design probes and detect those which have a high chance to crosshybridize with the genome. These are then discarded from the pool of possible sequence probes. The first step is to blast the sequence of the known gene with 1781 nucleotides against the entire yeast genome with 12,155,026 nucleotides. To reduce the computation time we take only a segment of 25 nucleotides out of gene and blasted it against the genome. All sequences with a E-value above a certain threshold are listed in table 2. In addition the segment is chosen from the 3'-region of the gene. This has the advantage that only complete mRNA will be detected because the mRNA is always synthesised from the 5' to 3' end. Then all free energies are calculated between the duplexes formed by the gene and the genome segments (see table 2). The segments which show a very low free energy like a_7 and a_{10} , could crosshybridize very stable and thus

| Sequence | Length | Score | Name | Free Energy kcal/mol |
|--------------------------|--------|-------|----------|----------------------|
| attagccgcttgggacgtcgagaa | 25 | 25 | segment | |
| cgcttgggacgt | 12 | 12 | a_1 | -12.10 |
| tagccgcttgg | 11 | 11 | a_2 | -11.03 |
| attagccgctt | 11 | 11 | a_3 | -9.66 |
| tgggacgtcgc | 11 | 11 | a_4 | -12.20 |
| acgtcgagaa | 11 | 11 | a_5 | -10.87 |
| cttgggacgtc | 11 | 11 | a_6 | -10.22 |
| gccgcttggga | 11 | 11 | a_7 | -12.30 |
| attagccgctt | 11 | 11 | a_8 | -9.66 |
| ggacgtcgcag | 11 | 11 | a_9 | -11.65 |
| tgggacgtcgc | 11 | 11 | a_{10} | -12.23 |

Table 2: Results of free energy calculation

should not be taken to detect the gene SE01. Whereas probes from the segments a_3 and a_8 are likely to form very stable crosshybridization and are probably ideal as probes.

5 Discussion

As anticipated the order of the different segments changes after calculating the free energy which shows that this two approaches BLASTn and NN-Thermodynamics classify the segments differently (see table 3). But for a more statis-

have a DNA solution with known concentration of one gene and measure the latter with two DNA-Chips with different probes. One probe would be the result of our algorithm generated for this gene, the other would be the outcome of another program. By comparing the resulting concentration of the two DNA-Chips with the real ones, we could judge the quality of the two probes and their sensitivity to crosshybridization.

| BLASTn | NN-Therm. | Free Energy kcal/mol |
|----------|-----------|----------------------|
| a_1 | a_7 | -12.3 |
| a_2 | a_{10} | -12.23 |
| a_3 | a_4 | -12.2 |
| a_4 | a_1 | -12.1 |
| a_5 | a_9 | -11.65 |
| a_6 | a_2 | -11.03 |
| a_7 | a_5 | -10.87 |
| a_8 | a_6 | -10.22 |
| a_9 | a_3 | -9.66 |
| a_{10} | a_8 | -9.66 |

Table 3: Comparison of the order of the segments generated with BLASTn and NN-Thermodynamics

tical foundation further calculations should be made. Another task would be the implementation of the proposed algorithm into a programme, which could perform the tasks automatically. Furthermore it would be interesting to compare the quality of the signal of the probes generated with our proposed algorithm and with those based on sequence databases and on thermodynamics (see section 2). In this scenario we would

| Software name | General Database search | | | | | | | | | | Further Characteristics | | | | Other parameters | |
|--------------------------------------|-----------------------------|---|------------------------------|------------------|-------------------------------|---------------------|---|--|-------------|---------------|------------------------------------|--------|---------------------------|----------------|------------------|--|
| | Procedure (BLASTn or FastA) | Input parameters | Minimum complementary length | Mismatch allowed | Position-dependent evaluation | E-Values, Identity | Further parameters | Result file | free-energy | Melting Temp. | Sec. structures | Others | experimental verification | 25Mer or 50Mer | | |
| Crosshybridization, Kachalo S. et al | no | Con. (target RNA), Signal intensity | no | no | no | no | - | Binding coefficient | no | no | no | - | 30Mer-60Mer | 8 Nucleotide | | |
| ChipCheck | no | free-energy (Target, Probe), Con. (Target), Probe | no | no | no | no | - | hybridization equilibria, free Enthalpie | no | no | no | - | 25Mer | - | | |
| PROBEMER | Suffix array search | Seq. (FASTA), 13 Databases | - | can be set | starting area can be set | Suffix array search | - | HTML List | yes | no | GC%, self-annealing, hairpin loops | yes | 25Mer | PRIMER3 | | |
| PRIMGENS | BLASTn, global alignment | Seq. (FASTA) | 50 Nt | 75% | no | yes/yes | - | - | yes | no | GC%, self-annealing, hairpin loops | yes | 25Mer | PRIMER3 | | |
| PROBEWIZ | BLASTn | Seq. (FASTA), Database 8 species | Default 20 Nt | % can be set | yes | yes/yes | Paralogy, Penalty, PRIMER3 | - | yes | no | GC%, self-annealing, hairpin loops | - | cDNA | PRIMER3 | | |
| OligoWiz | BLASTn | Seq. (FASTA) | Default 20 Nt | 75% | yes | yes/yes | Max. length cutoff | No Alignment output, score between 0 and 1 | yes | no | - | - | 20-30Mer or 70Mer | - | | |
| Sarani | similar to BLASTn | Seq. (FASTA) | - | - | no | no | - | - | yes | no | yes | - | 25Mer or cDNA | - | | |
| Oligodb | BLASTn | Seq. (FASTA) | - | - | no | yes/yes | - | HTML / tab delimited text file | yes | yes | low complexity region | no | oligos | - | | |
| Oligo Picker | BLASTn | Seq. (FASTA) | 10Nt | 100% Identity | yes | yes/yes | T _m , Probe copy, crosshybridization | - | yes | yes | cross reactivity, low complexity | - | 25Mer-100Mer | - | | |
| Oligo Array | BLASTn | Seq. (FASTA) | 50-15Nt | 50-100% Identity | yes | yes/yes | - | - | yes | yes | yes | - | oligos | - | | |

Table 4: Analysis of programmes for designing microarrays

Acknowledgment

We thank Dr. Moritz Diehl of the member of the Institut fuer Wissenschaftliches Rechnen (IWR) Heidelberg, whom lead the practical course "Mathematical Methods in Bioinformatics" in which this work emerged. I especially thank Barmak Mostofian and Kamila Naxerova who reviewed the document and submitted comments. Furthermore we want to thank Dr. Torsten Kroll from the Klinikum der Friedrich-Schiller-Universitaet Jena for stimulating discussion.

References

- [1] Benjamin Cummings, Addison Wesley Longman 2001
- [2] Affymetrix Manual (2001). Affymetrix Microarray Suite User Guide version 5.0. Santa Clara, CA.
- [3] Michael Baum et al., (2003) Validation of a novel, fully integrated and flexible microarray bench top facility for gene expression profiling. *Nucleic Acids Res.* **31**, 151
- [4] Stefanie B. Fulmer-Smentek, Ph.D. (2003). Performance comparison of Agilent's 60-mer and 25-mer *in situ* synthesised oligonucleotide microarrays. Agilent Manual
- [5] Henrik Bjorn Nielsen, Rasmus Wemerson and Steen Knudsen,(2003) Design of oligonucleotides for microarrays and perspectives for design of multi-transcriptional arrays. *Nucleic Acids Res.* **31**, 3491-3496
- [6] Jean-Marie et al., (2002) Oligo Array: genome-scale oligonucleotide design for microarrays. *Bioinformatics* **18**, 486-487
- [7] Dong Xu et al., (2002) PRIMEGENS: robust and efficient design of gene-specific probes for microarrays. *Bioinformatics* **18**, 1432-1437
- [8] Michael D. Kane et al., (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res.* **28**, 4542-4557
- [9] H. B. Nielsen et al., (2002) Avoiding crosshybridization by choosing non redundant targets on cDNA arrays. *Bioinformatics* **18**, 321-322
- [10] Xiaowei Wang and Brian Seed, (2003) Selection of Oligonucleotide Probes for Protein Coding Sequences. *Bioinformatics* **7**, 796-802
- [11] Ralf Mrowka et al., (2002) Oligodb-interactive design of oligo DNA for transcription profiling of human genes. *Bioinformatics* **18**, 1686-1687
- [12] Scott J. Emrich et al., (2003) PROBEmer: a web-based software tool for selecting optimal DNA oligos. *Nucleic Acids Res.* **31**, 3746-3750
- [13] Karsten H. Siegmund, Ulrich E. Steiner and Clemens Richert, (2003) ChipCheck-A Programme Predicting Total Hybridization Equilibria for DNA Binding to Small Oligonucleotide Microarrays. *J Chem Inf Comput Sci.* **43**(6), 2153-2162
- [14] SantaLucia et al., (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbour thermodynamics. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 1460-31465
- [15] Crothers, D. M. and Zimm, B. H. (1964) *J. Mol. Biol.* **9**, 403-410
- [16] Sarani, www.strandsgenomics.com
- [17] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic Local Alignment Search Tool. *J. Mol. Biol.* **215**, 403-410
- [18] Breslauer, K. J., Frank, R., Blocker, H., & MArky, L.A. (1986) *Proc. Natl. Acad. U.S.A.* **83**, 3746-3750
- [19] Borer, P.N., Dengler, B., Tinoco, I., Jr., & Uhlenbeck, O.C. (1974) *J. Mol. Biol.* **86**, 843-853