ADVANCES
IN BIOCHEMICAL ENGINEERING
BIOTECHNOLOGY

77

Managing Editor
T. Scheper

Volume Editor
J. Hoheisel

Chip Technology

Springer

77

Advances in Biochemical Engineering/Biotechnology

Managing Editor: T. Scheper

Advances in Biochemical Engineering/Biotechnology reviews actual trends in modern biotechnology. Its aim is to cover all aspects of this interdisciplinary technology where knowledge, methods and expertise are required for chemistry, biochemistry, microbiology, genetics, chemical engineering and computer science. Special volumes are dedicated to selected topics which focus on new biotechnological products and new processes for their synthesis and purification. They give the state-of-the-art of a topic in a comprehensive way thus being a valuable source for the next 3–5 years. It also discusses new discoveries and applications.

In general, special volumes are edited by well known guest editors. The managing editor and publisher will however always be pleased to receive suggestions and supplementary information. Manuscripts are accepted in English.

In references Advances in Biochemical Engineering/Biotechnology is abbreviated as *Adv Biochem Engin/Biotechnol* as a journal.

Visit the ABE home page at http://link.springer.de/series/abe/
http://link.springer-ny.com/series/abe/

ISSN 0724-6145
ISBN 3-540-43215-9
Springer-Verlag Berlin Heidelberg New York

Library of Congress Catalog Card Number 72-152360

Springer-Verlag Berlin Heidelberg New York
a member of BertelsmannSpringer Science+Business Media GmbH

http://www.springer.de

Printed in Germany

Typesetting: Fotosatz-Service Köhler GmbH, Würzburg
Cover: E. Kirchner, Heidelberg

Printed on acid-free paper SPIN: 10846806 02/3020mh – 5 4 3 2 1 0

Springer

Berlin
Heidelberg
New York
Barcelona
Hong Kong
London
Milan
Paris
Tokyo

Chip Technology

Volume Editor: Jörg Hoheisel

With contributions by
A. Brazma, K. Büssow, C. R. Cantor, F. C. Christians,
G. Chui, R. Diaz, R. Drmanac, S. Drmanac, H. Eickhoff,
K. Fellenberg, S. Hannenhalli, J. Hoheisel, A. Hou,
E. Hubbell, H. Jin, P. Jin, C. Jurinke, Z. Konthur,
H. Köster, S. Kwon, S. Lacy, H. Lehrach, R. Lipshutz,
D. Little, A. Lueking, G. H. McGall, B. Moeur,
E. Nordhoff, L. Nyarsik, P. A. Pevzner, A. Robinson,
U. Sarkans, J. Shafto, M. Sohail, E. M. Southern,
D. Swanson, T. Ukrainczyk, D. van den Boom, J. Vilo,
M. Vingron, G. Walter, C. Xu

Springer

Managing Editor

Professor Dr. T. Scheper
Institute of Technical Chemistry
University of Hannover
Callinstraße 3
30167 Hannover, Germany
E-mail: scheper@mbox.iftc.uni-hannover.de

Volume Editor

Dr. Jörg Hoheisel

Functional Genome Analysis
Deutsches Krebsforschungszentrum
Im Neuenheimer Feld 506
69120 Heidelberg
Germany
E-mail: j.hoheisel@dkfz-heidelberg.de

Editorial Board

Advances in Biochemical Engineering Biotechnology also Available Electronically

For all customers with a standing order for Advances in Biochemical Engineering/Biotechnology we offer the electronic form via LINK free of charge. Please contact your librarian who can receive a password for free access to the full articles. By registration at:

http://www.springer.de/series/abe/reg_form.htm

If you do not have a standard order you can nevertheless browse through the table of contents of the volumes and the abstracts of each article at:

http://link.springer.de/series/abe/
http://link.springer_ny.com/series/abe/

There you will find also information about the

- Editorial Board
- Aims and Scope
- Instructions for Author

Attention all Users of the "Springer Handbook of Enzymes"

Information on this handbook can be found on the internet at
http://www.springer.de/enzymes/

A complete list of all enzymes entries either as an alphabetical Name Index or as the EC-Number Index is available at the above mentioned URL. You can download and print them free of charge.

A complete list of all synonyms (more than 25,000 entries) used for the enyzmes is available in print form, ISBN 3-540-41830-X.

Save 15%

We recommend a standing order for the series to ensure you automatically receive all volumes and all supplements and save 15% on the list price.

Preface

DNA-chip analysis has come a long way since the first official gathering of scientists working in this field, which took place in Moscow in 1991, comprising about 50 scientists from Europe and the USA. Then, the initial aim was the development of a technology for high-throughput sequencing of the human genome, named sequencing by hybridisation. The field soon diversified, however, in terms of methodology and application. Nowadays, DNA-microarrays seem to be a common commodity in biological sciences. The complexity hidden behind the apparent ease of such studies, however, is highlighted by the fact that it took about ten years of technology development – and persuasion – before the methodology really took off. Also, at closer scrutiny one realises that many problems still need to be resolved and only relatively limited inquiries have been attempted so far considering the potentially possible. Nevertheless, even these experiments have produced data on a scale beyond imagination for most people in biology a few years ago and for many even today. Even the data sets originating from large-scale sequencing are dwarfed by the quantity of information from chip-based analyses. Thus, DNA-microarray technology will be the true herald of change in biology. Having developed earlier from a descriptive to an analytical science, biology will split in future into an experimental and a theoretical branch, similar to what happened in physics quite a long time ago.

This change in biology is no better represented than by the authors of this book, who took part in bringing about this shift in emphasis. They are well-known experts in the field, many – like Edwin Southern, Hans Lehrach, Radoje Drmanac, Pavel Pevzner and Charles Cantor – have been actively pursuing array technology for more than a decade. Rather than pondering on the good old times, however, they demonstrate the continuous development in both technology and application areas and elucidate some critical points that need to be considered, when performing microarray analyses.

The first article, by Pavel Pevzner and co-workers, informs on aspects of designing DNA-arrays, which is not a trivial undertaking, although rarely discussed. Even at this level of array-based analysis – right at the start rather than the end – informatics is required in order to deal with the inherent complexity.

Such a design phase is followed by the actual production process. Although by no means the sole procedure for doing so, photolithographically controlled in situ synthesis is currently the best advanced technique for large-scale chip production with a high degree of reproducibility. Glen McGall and Fred Christians report on the procedures involved, some future developments and applications.

Muhammad Sohail and Edwin Southern describe in their contribution a different process for the creation of oligonucleotide arrays. Subsequently, they used the arrays in the screening for effective antisense reagents. This study is fascinating because of its impact on the understanding of interactions between nucleic acids, an interesting research area even after 50 years of structural DNA analysis, and in view of its practical implications for applications in biotechnology and molecular medicine.

Hubert Köster, Charles Cantor and co-workers combine two high-throughput techniques – arrays with detection by mass spectrometry – for genotyping analyses. With the existence of about 1.5 million publicly known single nucleotide polymorphisms (SNPs), the exploration of this resource is a major challenge in the effort of translating basic genomics into applied, medically relevant molecular genetics.

The work of Radoje Drmanac extends the objective of genotyping analyses and, at the same time, returns to the origin of DNA-chip analysis by pursuing 'sequencing by hybridisation', which is nothing short of performing a SNP analysis for each and every nucleotide of a given DNA-fragment. He reports recent achievements and deliberates on the exciting opportunities of this methodology.

The text of Holger Eickhoff and colleagues already reaches beyond the mere DNA-chip by reporting on procedures that extend strongly into the field of proteomics, thus linking the two areas. Only by such measures, carried out experimentally as well as in silico, the complexity of functions in cellular systems will eventually be unravelled.

Considering how array analyses are performed, it is only natural that a contribution on bioinformatics tools and databases should come at the end of the list. Its position does not reflect its importance, however. As a matter of fact, all preparatory work going into the production of the nice looking, colourful pictures from DNA-arrays is a useless squander unless it is assessed and presented in a way that makes the data accessible to human interpretation. Currently, much of the data produced on microarrays is actually wasted. Transcriptional profiling studies, for example, usually concentrate on few, specific biological aspects and ignore much else contained in the very set of raw data. This information could be useful for other studies, if only one could access it. For the purpose of going back to results for entirely different analytical purposes, central databases with appropriately designed and standardised procedures, as well as a common ontology, are essential. Alvis Brazma and colleagues have been instrumental in getting such efforts started.

Overall, the various articles provide a good mix, covering many, although not all, aspects of microarray-based analysis, the latter no longer achievable in a single book, for the days of the Moscow meeting in 1991 are long past and the breadth of the field has expanded enormously both in terms of its technical aspects and the variety of potential applications. Nevertheless, I hope the picture is comprehensive enough for understanding the basics, elaborate enough to inform in detail on certain aspects and speculative enough to stimulate further developments.

Heidelberg, May 2002 Jörg D. Hoheisel

Contents

Combinatorial Algorithms for Design of DNA Arrays

Sridhar Hannenhalli[1] · Earl Hubbell[2] · Robert Lipshutz[2] · Pavel A. Pevzner[3]

[1] Departments of Mathematics, University of Southern California, Los Angeles, California 90089-1113, USA. Present address: Celera Genomics, Rockville, Maryland 20850, USA
[2] Affymetrix, 3380 Central Expressway Santa Clara, California 95051, USA. *E-mail: earl_Hubbell@affymetrix.com*
[3] Departments of Computer Science and Engineering, University of California at San Diego, La Jolla, California 92093-0114, USA. *E-mail: ppevzner@cs.ucsd.edu*

Optimal design of DNA arrays requires the development of algorithms with two-fold goals: reducing the effects caused by unintended illumination (*border length minimization problem*) and reducing the complexity of masks (*mask decomposition problem*). We describe algorithms that reduce the number of rectangles in mask decomposition by 20–30% as compared to a standard array design under the assumption that the *arrangement* of oligonucleotides on the array is fixed. This algorithm produces provably optimal solution for all studied real instances of array design. We also address the difficult problem of finding an arrangement which minimizes the border length and come up with a new idea of *threading* that significantly reduces the border length as compared to standard designs.

Keywords: DNA arrays, Photolitography, Mask design, Combinatorial algorithms, Rectangle cover problem

Advances in Biochemical Engineering/
Biotechnology, Vol. 77
Managing Editor: T. Scheper

1 Introduction

DNA arrays have been greatly successful in a variety of genomic analyses [2], ranging from detecting SNPs [13] to functional genomics [3, 9, 14]. One method of synthesizing an oligonucleotide array is by a photolithographic *V LSIPS* (*very large scale immobilized polymer synthesis*) method similar to that used in the semiconductor industry. In this method, light is used to direct synthesis of oligonucleotides in an array. In each step, light is selectively allowed through a mask to expose cells in the array, activating the oligonucleotides in that cell for further synthesis. For every synthesis step there is a *mask* with corresponding open (allowing light) and closed (blocking light) cells. Each mask corresponds to a step of combinatorial synthesis described in [5].

The proper regions are activated by illuminating the array through a series of masks. Unfortunately, because of diffraction, internal reflection, and scattering, points close to the border between an illuminated region and a shadowed region are often subject to unintended illumination [5]. In such a region, it is uncertain whether a nucleotide will be appended or not. This uncertainty gives rise to probes with unknown sequences complicating interpretation of the experimental data. Methods are being sought to minimize the lengths of these borders so that the level of uncertainty is reduced. The problem is important not only for Affymetrix arrays but also for any other in-situ synthesis scheme, such as the recently published micromirror arrays [12].

Border Minimization Problem (BMP) is to arrange the probes on the array in such a way that the overall border length of *all* masks is minimal. For two probes x and y, let $\delta(x, y)$ be the Hamming distance between probes, i. e., the number of positions in which x and y differ. An intuitive approach to BMP is to position similar probes at neighboring sites. In a *synchronous* array design when every iteration generates a fixed nucleotide at i-th positions of the probes, the overall border length of all masks equals $\sum \delta(x, y)$, where the sum is taken over all pairs of neighboring probes on the array. For uniform arrays this observation establishes the connection between minimization of border length and Gray codes [11]. An l-bit *Gray* code is defined as a permutation of the binary numbers between 0 and $2^l - 1$ such that neighboring numbers have exactly one differing bit, as do the first and last numbers.

For a fixed set of n probes, consider an n-vertex complete graph with edge weights equal to Hamming distances between probes. For uniform arrays with 4^l probes, every solution of the *Traveling Salesman Problem (TSP)* corresponds to a Gray code and leads to arrangements that position similar probes in neighboring sites thus minimizing the border length of the mask for uniform arrays [11]. However, for *assay* arrays with arbitrary probes used in functional genomics, the border length minimization remained an open proplem.

The goal of positioning similar probes in neighboring sites led to the attempts of using TSP approach for border length minimization in the design of dedicated functional genomics arrays. In the very first design of such arrays at Affymetrix in mid 1990s Earl Hubbell implemented a greedy TSP algorithm for minimizing the border length of the assay chip. The motivation for using a TSP tour is that

consecutive probes in the tour are likely to be similar. The probe arrangement corresponding to found TSP solution was further *threaded* on an array in a row-by-row fashion. The deficiency of this approach is that although probes in the same row are usually similar, the neighboring probes in the same column may be very different thus leading to an increase in border length. In this paper we propose a different idea for array design called *optimal threading*. Define a threading as a self-avoiding path through the sites of an array. Different threadings vary in the border length and we show that a clever choice of threading reduces the border length by 20% for large chips (as compared to the row-by-row threading).

Masks for DNA arrays are built by patterning equipment which generates rectangular regions by a long series of "flashes", each flash producing a rectangle. Complex masks are made by consecutive generation of rectangles. The cost of a mask is (approximately) linear in the number of rectangles generated, and there is a fixed upper limit to the number of rectangles it is feasible to generate, before a different (expensive) mask fabrication technology must be used instead. Therefore, we are interested in a cover of a mask with a minimum number of rectangles (the rectangles in a cover can overlap).

The rectangle cover problem is known to be NP-hard and approximation algorithms with small performance ratios for this problem are unknown ([6, 7]). We explore the specifics of oligonucleotide masks and devise an efficient algorithm which finds a (provably) optimal rectangle cover for all masks we tested at Affymetrix.

The paper is organized as follows. In the following section we present the "Placement problem" which is a generalization of a few interesting optimization problem, including optimal mask design and (surprisingly) protein folding in the lattice model. We further describe applications of TSP for DNA array design and present threading as novel approach to optimizing the border length. Finally, we describe the algorithms for optimal mask decomposition into rectangles.

2 Placement Problem

Let $G_1(V_1, E_1, w_1)$ and $G_2(V_2, E_2, w_2)$ be two complete edge-weighted graphs with weight functions w_1 and w_2. A bijective function $\psi: V_2 \to V_1$ is called a *placement* of G_2 on G_1. The cost of placement is defined as

$$c(\psi) = \sum_{x, y \in V_2} w_1(x, y)\, w_2(\psi(x), \psi(y)).$$

The *optimal placement* problem is to find a placement of G_2 on G_1 of minimal cost.

For optimal mask design problem the $N \times N$ vertices of the *grid* graph G_1 correspond to a $N \times N$ array. The distance between grid cells (i_1, j_1) and (i_2, j_2) in the *Manhattan metric* is defined as $|i_2 - i_1| + |j_2 - j_1|$. The vertices of the grid are *neighbors* if the distance between them is 1. The weight $w_1(x, y)$ is defined as 1 if x and y are neighbors and 0 otherwise. The vertices of the *probe* graph G_2 correspond to $N \times N$ probes to be synthesized on a chip. The weight function $w_2(p, q)$ is de-

Table 1. Performance comparison of various algorithms for real masks

Chip Size	549a8 312×312	549a14 312×312	549a18 312×312	549a19 312×312	549a22 312×312
#Black Cells	17377	11152	14279	32519	16102
#Maximal Rectangles	1394	993	2362	2952	3425
#Prime Rectangles	1392	968	2327	2920	3387
#Black Cells after Phase 1	4	260	610	509	433
#iterations in Phase 2	1	1	1	1	1
#Rectangles found in Phase 2	2	17	35	32	34
#Rectangles found after Phase 2	0	0	0	0	0
#Rectangles found by Rectangle_Cover	1394	985	2362	2952	3421
#Rectangles found by Rectangle_Partition	1394	985	2507	3406	3521
#Rectangles found by Rectangle_Greedy	1513	1432	2980	4588	3816

fined as the *distance* between probes p and q. The distance between probes is the number of times an *edge* occurs between them in the entire set of masks to fabricate the chip. Depending on the synthesis strategy the distance is either Hamming distance or the distance defined by the synthesis-schedule. In such formulation the optimal placement attempts to position similar probes in the neighboring position of the grid and minimizes the *overall border length* of the masks. In many cases the overall border length is well-correlated with the number of flashes to produce the mask and therefore, the optimal placement reduces the cost of the mask set in these cases (Table 1).

TSP is a particular case of the optimal placement problem (in this case G_1 is a graph with edges of weight 1 forming a path and all other edges of weight 0). This observation motivates the following approach to optimal mask design problem: find TSP in the probe graph and *thread* it in a row-by-row fashion on the array grid. Such approach optimizes the placement for half of the edges (for adjacent cells in the same row) in the grid but completely ignores another half (for adjacent cells in the same column). Experiments suggest that this approach reduces the cost of the placement by 27% as compared to the random design for 64 Kb chip. The statistical analysis indicates that the larger is the chip the larger is the saving in cost (four-fold increase in the size of the chip increase the saving by about 5%). Our goal is to optimize the placement for the remaining half of the edges and to reduce the cost of the edges by additional 27%. Below we show how to reduce it by 20% thus achieving savings which is close to optimal.

3 DNA Arrays and the Traveling Salesman Problem

Since any improvement of the TSP-tour would likely improve the parameters of array design, we first tried to find a better TSP algorithm for assay arrays. Computing an optimal TSP tour for a graph with 64,000 vertices is not feasible. Note that only very fast approximation algorithms (i. e. at most quadratic) are suitable for our size of the problem. We implemented 2-opt algorithm [8] for TSP. Given an ordering of vertices $\pi = \pi_1 \dots \pi_{i-1}, \pi_i, \pi_{i+1} \dots \pi_{j-1}, \pi_j, \pi_{j+1} \dots \pi_n$, a *reversal* of π at i, j is a permutation $\pi' = \pi_1 \dots \pi_{i-1}, \pi_j, \pi_{j-1} \dots \pi_{i+1}, \pi_i, \pi_{j+1} \dots \pi_n$ obtained from π by reversing the fragment $\pi_i, \pi_{i+1} \dots \pi_{j-1}, \pi_j$. *2-opt neighborhood* of π is formed by all permutations obtained from π by reversals. The Lin-Kernighan local improvement algorithm finds a suboptimal TSP tour by starting from an arbitrary permutation of vertices $\pi = \pi(0)$. At a general step, given a permutation $\pi(i)$, it searches for a permutation $\pi(i+1)$ from 2-opt neighborhood of $\pi(i)$ such that the cost of $\pi(i+1)$ is less than the cost of $\pi(i)$. The computation terminates at a local minimum when no such permutation in 2-opt neighborhood exists.

Disappointingly, we found that a 2-opt optimization of the greedy TSP-tour leads to only 0.3% improvement (Table 1). We tried a few approaches to further improve the TSP solution (or to verify its proximity to the optimal solution), in particular, solving *minimum length cycle cover* problem with further merging of cycles in the cover into a hamiltonian path. A cycle cover is a set of (simple) cycles such that each vertex in the graph is contained in exactly one cycle. The

Table 2. Performance comparison of various algorithms for random masks with varying density; the parameter p denotes that probability of a grid cell being a black cell

Chip Size	r(p=0.9) 500×500	r(p=0.75) 500×500	r(p=0.5) 500×500	r(p=0.25) 500×500	r(p=0.1) 500×500
#Black Cells	224660	186891	123951	60791	24241
#Maximal Rectangles	50944	65399	61364	41092	20362
#Prime Rectangles	8452	28608	48399	39627	20322
#Black Cells after Phase 1	97522	37771	6400	339	24241
#iterations in Phase 2	7	11	5	2	1
#Rectangles found in Phase 2	1343	9599	3513	272	1
#Rectangles found after Phase 2	12561	2393	18	0	0
#Rectangles found by Rectangle_Cover	22356	40600	51930	39899	20323
#Rectangles found by Rectangle_Partition	22726	43633	53123	39977	20323
#Rectangles found by Rectangle_Greedy	28420	45900	55077	40267	20337