Guide to Analysis of DNA Microarray Data

SECOND EDITION

Steen Knudsen

Guide to ANALYSIS OF DNA MICROARRAY DATA

Second Edition

Steen Knudsen
Center for Biological Sequence Analysis
BioCentrum-DTU

Technical University of Denmark

WILEY-LISS

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data is available.

ISBN 0-471-65604-6

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

To Linnea

Preface

I am often asked, "Do you have a good text I can read on analysis of DNA array data?" This is an attempt at providing such a text for students and scientists alike who venture into the field of DNA array data analysis for the first time. The book is written for biologists and medical researchers without special training in data analysis and statistics. Mathematical stringency is sacrificed for intuitive and visual introduction of concepts. Methods are introduced by simple examples and citations of relevant literature. Practical computer solutions to common analysis problems are suggested, with an emphasis on software developed at and made freely available by my own lab. The text emphasizes gene expression analysis.

This text takes over where the DNA array equipment leaves you: with a file containing an image of the microarray. If the equipment has already performed an analysis of the image, you are left with a file of signal intensities. The information in that file will prompt questions such as: How is it scaled? What is the error in the data? When can I say that a certain gene is up-regulated? What do I do with the thousands of genes that show some regulation? How much information can I get out of my data? This text will attempt to answer those questions and others that will come into mind as you delve further into the data.

Since the appearance of the first edition, the field has virtually exploded, with thousands of papers published on DNA microarrays and data analysis. A new generation of microarray equipment, allowing *in situ* synthesis of chips, has appeared. New public software packages have appeared, and improved

methods for data analysis have been published. The second edition includes all these new and recent developments and also contains new chapters on image analysis, experiment design, interpretation of results, oligonucleotide probe design, data integration, and systems biology. The second edition aims to be the most comprehensive and up-to-date book available on DNA microarrays.

Each chapter has a section on Further Reading, which categorizes key literature by topic.

A web companion site[1] is available with copy-paste code examples from the book, errata, experimental protocols, and more.

STEEN KNUDSEN

Lyngby, Denmark
December 2003

[1] http://www.cbs.dtu.dk/steen/book.html

Acknowledgments

Christopher Workman, Laurent Gautier, and Henrik Bjørn Nielsen inspired me for many aspects of this book and also implemented many methods used in the book.

I thank Yves Moreau for helpful suggestions on the manuscript.

I thank my collaborators Claus Nielsen, Kenneth Thirstrup, Torben Ørntoft, Friedrik Wikman, Thomas Thykjaer, Mogens Kruhøffer, Karin Demtröder, Hans Wolf, Lars Dyrskjøt Andersen, Casper Møller Frederiksen, Jeppe Spicker, Lars Juhl Jensen, Carsten Friis, Hanne Jarmer, Hans-Henrik Saxild, Randy Berka, Matthew Piper, Steen Westergaard, Christoffer Bro, Thomas Jensen, and Kristine Dahlin for allowing me to use examples generated from our collaborative research.

I am grateful to Center Director Søren Brunak for creating the environment, and to the Danish National Research Foundation, the Danish Biotechnology Instrument Center, and Novozymes A/S for funding the research that made this book possible.

S. K.

Contents

1

Introduction to DNA Microarray Technology

1.1 HYBRIDIZATION

The fundamental basis of DNA microarrays is the process of *hybridization*. Two DNA strands hybridize if they are complementary to each other. Complementarity reflects the Watson-Crick rule that adenine (A) binds to thymine (T) and cytosine (C) binds to guanine (G). One or both strands of the DNA hybrid can be replaced by RNA and hybridization will still occur as long as there is complementarity.

Hybridization has for decades been used in molecular biology as the basis for such techniques as Southern blotting and Northern blotting. In Southern blotting, a small string of DNA, an *oligonucleotide*, is used to hybridize to complementary fragments of DNA that have been separated according to size in a gel electrophoresis. If the oligonucleotide is radioactively labeled, the hybridization can be visualized on a photographic film that is sensitive to radiation. In Northern blotting, a radio-labeled oligonucleotide is used to hybridize to messenger RNA that has been run through a gel. If the oligo is specific to a single messenger RNA, then it will bind to the location (*band*) of that messenger in the gel. The amount of radiation captured on a photographic film depends to some extent on the amount of radio-labeled probe present in the band, which again depends on the amount of messenger. So this method is a semiquantitative detection of individual messengers.

DNA arrays are a massively parallel version of Northern and Southern blotting. Instead of distributing the oligonucleotide probes over a gel containing samples of RNA or DNA, the oligonucleotide probes are attached

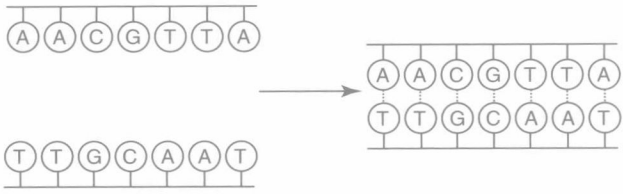

Fig. 1.1 Hybridization of two DNA molecules. Dotted line: hydrogen bonds.

to a surface. Different probes can be attached within micrometers of each other, so it is possible to place many of them on a small surface of one square centimeter, forming a DNA array. The sample is labeled fluorescently and added to the array. After washing away excess unhybridized material, the hybridized material is excited by a laser and is detected by a light scanner that scans the surface of the chip. Because you know the location of each oligonucleotide probe, you can quantify the amount of sample hybridized to it from the image generated by the scan.

There is some contention in the literature on the use of the word "probe" in relation to microarrays. Throughout this book the word "probe" will be used to refer to what is attached to the microarray surface. And the word "target" will be used to refer to what is hybridized to the probes.

Where before it was possible to run a couple of Northern blots or a couple of Southern blots in a day, it is now possible with DNA arrays to run hybridizations for tens of thousands of probes. This has in some sense revolutionized molecular biology and medicine. Instead of studying one gene and one messenger at a time, experimentalists are now studying many genes and many messengers at the same time. In fact, DNA arrays are often used to study *all* known messengers of an organism. This has opened the possibility of an entirely new, systemic view of how cells react in response to certain stimuli. It is also an entirely new way to study human disease by viewing how it affects the expression of all genes inside the cell. Figure 1.2 illustrates the revolution of DNA arrays in biology and medicine by the number of papers published on the topic.

1.2 GOLD RUSH?

The explosion in interest in DNA microarrays has almost been like a gold rush. Is there really that much gold to be found with this new technology? I am afraid that, in the short term, there will be some disappointments. Yes, you can learn about the gene expression in your organism or disease of interest,

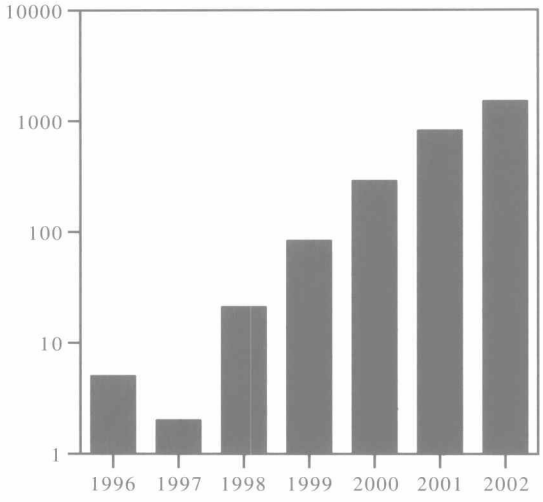

Fig. 1.2 The number of papers published per year referring to DNA microarrays.

but does that make you wiser? Typically, the wealth of data generated results in more questions than answers. There is one exception to this, and that is where DNA arrays have been used for diagnostics and prognostics. Here, DNA arrays have shown promising results in almost all the fields where they have been applied. This is where I think that the greatest short-term success of DNA microarray technology lies.

On a longer time scale molecular biology will benefit tremendously from the systemic approach offered by DNA microarrays and other massively parallel approaches. Many important discoveries lie in the interpretation of microarray data – more so from large compilations of experiments and large-scale experiments than from small experiments with just a few arrays.

1.3 THE TECHNOLOGY BEHIND DNA MICROARRAYS

When DNA microarrays are used for measuring the concentration of messenger RNA in living cells, a *probe* of one DNA strand that matches a particular messenger RNA in the cell is used. The concentration of a particular messenger is a result of *expression* of its corresponding gene, so this application is often referred to as *expression analysis*. When different probes matching all messenger RNAs in a cell are used, a snapshot of the total messenger RNA

pool of a living cell or tissue can be obtained. This is often referred to as an *expression profile* because it reflects the expression of every single measured gene at that particular moment. Expression profile is also sometimes used to describe the expression of a single gene over a number of conditions.

Expression analysis can also be performed by a method called *serial analysis of gene expression* (SAGE). Instead of using microarrays, SAGE relies on traditional DNA sequencing to identify and enumerate the messenger RNAs in a cell (see Section 1.3.6).

Another traditional application of DNA microarrays is to detect mutation in specific genes. The massively parallel nature of DNA microarrays allows the simultaneous screening of many, if not all, possible mutations within a single gene. This is referred to as *genotyping* (Chapter 12).

The treatment of array data does not depend so much on the technology used to gather the data as it depends on the application in question. Genotyping and expression analysis are two completely different applications, and they will be treated separately in this text. Most of the information will address analysis of expression data, and a separate chapter will address genotyping chips.

For expression analysis the field has been dominated in the past by two major technologies. The Affymetrix, Inc. GeneChip system uses prefabricated oligonucleotide chips (Figures 1.3 and 1.4). Custom-made chips use a robot to spot cDNA, oligonucleotides, or PCR products on a glass slide or membrane(Figure 1.5).

More recently, several new technologies have entered the market. In the following, several of the major technology platforms for gene expression analysis will be described.

1.3.1 Affymetrix GeneChip Technology

Affymetrix uses equipment similar to that which is used for making silicon chips for computers, and thus allows mass production of very large chips at reasonable cost. Where computer chips are made by creating masks that control a photolithographic process for removal or deposition of silicon material on the chip surface, Affymetrix uses masks to control synthesis of oligonucleotides on the surface of a chip. The standard phosphoramidite method for synthesis of oligonucleotides has been modified to allow light control of the individual steps. The masks control the synthesis of several hundred thousand squares, each containing many copies of an oligo. So the result is several hundred thousand different oligos, each of them present in millions of copies.

That large number of oligos, up to 25 nucleotides long, has turned out to be very useful as an experimental tool to replace all experimental detection procedures that in the past relied on using oligonuclotides: Southern, Northern, and dot blotting as well as sequence specific probing and mutation detection.

For expression analysis, up to 40 oligos are used for the detection of each gene. Affymetrix has chosen a region of each gene that (presumably) has the

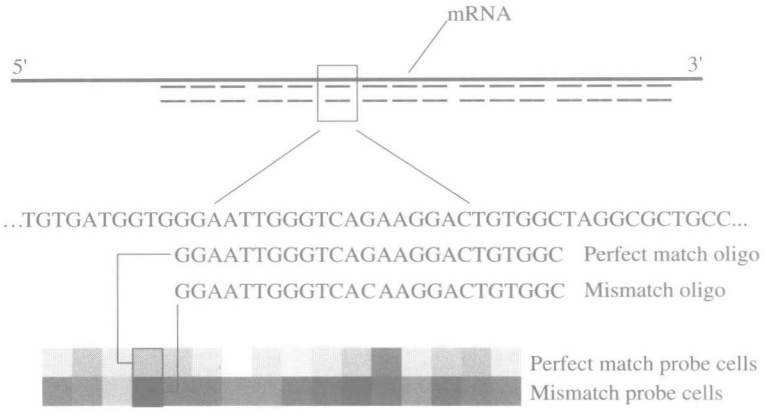

Fig. 1.3 The Affymetrix GeneChip technology. The presence of messenger RNA is detected by a series of probe pairs that differ in only one nucleotide. Hybridization of fluorescent messenger RNA to these probe pairs on the chip is detected by laser scanning of the chip surface. (Figure by Christoffer Bro.)

least similarity to other genes. From this region 11 to 20 oligos are chosen as perfect matches (PM) (i.e., perfectly complementary to the mRNA of that gene). In addition, they have generated 11 to 20 mismatch oligos (MM), which are identical to the PM oligos except for the central position 13, where one nucleotide has been changed to its complementary nucleotide. Affymetrix claims that the MM oligos will be able to detect nonspecific and background hybridization, which is important for quantifying weakly expressed mRNAs. However, for weakly expressed mRNAs where the signal-to-noise ratio is smallest, subtracting mismatch from perfect match adds considerably to the noise in the data (Schadt et al., 2000). That is because subtracting one noisy signal from another noisy signal yields a third signal with even more noise.

The hybridization of each oligo to its target depends on its sequence. All 11 to 20 PM oligos for each gene have a different sequence, so the hybridization will not be uniform. That is of limited consequence as long as we wish to detect only *changes* in mRNA concentration between experiments. How such a change is calculated from the intensities of the 22 to 40 probes for each gene will be covered in Section 4.3.

To detect hybridization of a target mRNA by a probe on the chip, we need to label the target mRNA with a fluorochrome. As shown in Figure 1.4, the steps from cell to chip usually are as follows:

- Extract total RNA from cell (usually using TRIzol from Invitrogen or RNeasy from QIAGEN).