Cancer Diagnostics with DNA Microarrays

Steen Knudsen
Medical Prognosis Institute, Hørsholm, Denmark
Cancer Diagnostics with DNA Microarrays
Cancer Diagnostics with DNA Microarrays

Steen Knudsen
Medical Prognosis Institute, Hørsholm, Denmark
To Tarja
Contents

Preface xv
Acknowledgments xvii

1. Introduction to DNA Microarray Technology 1
   1.1 Hybridization / 1
   1.2 The Technology Behind DNA Microarrays / 3
      1.2.1 Affymetrix GeneChip Technology / 4
      1.2.2 Spotted Arrays / 6
      1.2.3 Digital Micromirror Arrays / 8
      1.2.4 Inkjet Arrays / 8
      1.2.5 Bead Arrays / 9
      1.2.6 Electronic Microarrays / 9
      1.2.7 Serial Analysis of Gene Expression (SAGE) / 10
   1.3 Parallel Sequencing on Microbead Arrays / 11
      1.3.1 Emerging Technologies / 12
   1.4 Summary / 13
      Further Reading / 14

2. Image Analysis 16
   2.1 Gridding / 16
   2.2 Segmentation / 17
   2.3 Intensity Extraction / 17
   2.4 Background Correction / 17
CONTENTS

2.5 Software / 19
2.5.1 Free Software for Array Image Analysis / 19
2.5.2 Commercial Software for Array Image Analysis / 19
2.6 Summary / 20

3. Basic Data Analysis / 22
3.1 Normalization / 22
3.1.1 One or More Genes Assumed Expressed at Constant Rate / 23
3.1.2 Sum of Genes Is Assumed Constant / 24
3.1.3 Subset of Genes Is Assumed Constant / 24
3.1.4 Majority of Genes Assumed Constant / 24
3.1.5 Logit Normalization / 24
3.1.6 Spike Controls / 25
3.2 Dye Bias, Spatial Bias, Print Tip Bias / 25
3.3 Expression Indices / 26
3.3.1 Average Difference / 26
3.3.2 Signal / 26
3.3.3 Model-Based Expression Index / 26
3.3.4 Robust Multiarray Average / 27
3.3.5 Position-Dependent Nearest Neighbor Model / 27
3.3.6 Logit-t / 27
3.4 Detection of Outliers / 28
3.5 Fold Change / 28
3.6 Significance / 29
3.6.1 Multiple Conditions / 30
3.6.2 Nonparametric Tests / 30
3.6.3 Correction for Multiple Testing / 30
3.6.4 Example I: t-Test and ANOVA / 31
3.6.5 Example II: Number of Replicates / 32
3.7 Mixed Cell Populations / 33
3.8 Summary / 34
Further Reading / 34

4. Visualization by Reduction of Dimensionality / 37
4.1 Principal Component Analysis / 37
4.2 Supervised Principal Component Analysis / 39
4.3 Independent Component Analysis / 39
4.4 Example I: PCA on Small Data Matrix / 39
4.5 Example II: PCA on Real Data / 41
4.6 Summary / 42
Further Reading / 42

5. Cluster Analysis / 44
5.1 Hierarchical Clustering / 44
5.2  \textit{K}-Means Clustering / 46
5.3  Self-Organizing Maps / 46
5.4  Distance Measures / 47  
  \hspace{1em} 5.4.1  Example: Comparison of Distance Measures / 49
5.5  Time-Series Analysis / 51
5.6  Gene Normalization / 51
5.7  Visualization of Clusters / 52
5.8  Summary / 52
Further Reading / 52

6.  \textbf{Molecular Classifiers for Cancer}  \hspace{1em} 54

6.1  \textbf{Supervised versus Unsupervised Analysis} / 54
6.2  \textbf{Feature Selection} / 55
6.3  \textbf{Validation} / 55
6.4  \textbf{Classification Schemes} / 56  
  \hspace{1em} 6.4.1  \textit{Nearest Neighbor} / 56
  \hspace{1em} 6.4.2  \textit{Nearest Centroid} / 57
  \hspace{1em} 6.4.3  \textit{Neural Networks} / 57
  \hspace{1em} 6.4.4  \textit{Support Vector Machine} / 58
6.5  \textbf{Performance Evaluation} / 58
6.6  \textbf{Example I: Classification of SRBCT Cancer Subtypes} / 59
6.7  \textbf{A Network Approach to Molecular Classification} / 59  
  \hspace{1em} 6.7.1  \textit{Construction of HG-U95Av2-Chip Based Gene Network} / 61
  \hspace{1em} 6.7.2  \textit{Construction of HG-U133A-Chip Based Gene Network} / 61
  \hspace{1em} 6.7.3  \textit{Example: Lung Cancer Classification with Mapped Subnetwork} / 61
  \hspace{1em} 6.7.4  \textit{Example: Brain Cancer Classification with Mapped Subnetwork} / 62
  \hspace{1em} 6.7.5  \textit{Example: Breast Cancer Classification with Mapped Subnetwork} / 62
  \hspace{1em} 6.7.6  \textit{Example: Leukemia Classification with Mapped Subnetwork} / 62
6.8  \textbf{Summary} / 63
Further Reading / 64

7.  \textbf{Survival Analysis}  \hspace{1em} 66

7.1  \textit{Kaplan–Meier Approach} / 66
7.2  \textit{Cox Proportional Hazards Model} / 67
7.3  \textit{Software} / 68
Further Reading / 68

8.  \textbf{Meta-Analysis}  \hspace{1em} 69

8.1  \textit{Gene Matching} / 69
8.2  \textit{Combining Measurements} / 70  
  \hspace{1em} 8.2.1  \textit{Intensity} / 70
11.4 Microarray Studies of Breast Cancer / 90
   11.4.1 The Stanford Group / 90
   11.4.2 The Netherlands Cancer Institute Group / 91
   11.4.3 The Duke University Group / 92
   11.4.4 The Lund Group / 92
   11.4.5 The Karolinska Group / 93
   11.4.6 The Tokyo Group / 93
   11.4.7 The Veridex Group / 93
   11.4.8 The National Cancer Institute Group / 93
11.5 Meta-Classification of Breast Cancer / 93
11.6 Summary / 95
Further Reading / 95

12. Leukemia

12.1 Current Diagnosis / 98
12.2 Current Therapy / 98
12.3 Microarray Studies of Leukemia / 99
   12.3.1 The Boston Group / 99
   12.3.2 The Austria Group / 99
   12.3.3 The Utah Group / 99
   12.3.4 The Memphis Group / 100
   12.3.5 The Japan Group / 100
   12.3.6 The Munich Group / 100
   12.3.7 The Stanford Group / 101
   12.3.8 The Copenhagen Group / 101
   12.3.9 The Netherlands Group / 101
12.4 Summary / 101
Further Reading / 101

13. Lymphoma

13.1 Microarray Studies of Lymphoma / 105
   13.1.1 The Stanford Group / 105
   13.1.2 The Boston Group / 105
   13.1.3 The NIH Group / 105
   13.1.4 The NCI Group / 106
13.2 Meta-Classification of Lymphoma / 106
   13.2.1 Matching of Components / 107
13.2.2 Blind Test Set / 109
13.3 Summary / 112
Further Reading / 112

14. Lung Cancer

14.1 Microarray Studies of Lung Cancer / 114
   14.1.1 The Harvard/Whitehead Group / 114
14.1.2 The Minnesota Group / 116
14.1.3 The Vanderbilt Group / 116
14.1.4 The Tokyo Group / 116
14.1.5 The Michigan Group / 116
14.1.6 The Mayo Clinic / 116
14.1.7 The Toronto Group / 117
14.1.8 The NIH Group / 117
14.1.9 The Stanford Group / 117
14.1.10 The Israel Group / 117
14.2 Meta-Classification of Lung Cancer / 117
14.3 Summary / 122
Further Reading / 122

15. Bladder Cancer

15.1 Microarray Studies of Bladder Cancer / 124
  15.1.1 The Aarhus Group / 124
  15.1.2 The New York Group / 125
  15.1.3 The Dusseldorf Group / 125
15.2 Summary / 125
Further Reading / 125

16. Colon Cancer

16.1 Microarray Studies of Colon Cancer / 127
  16.1.1 The Princeton Group / 127
  16.1.2 The Maryland Group / 128
  16.1.3 The Aarhus Group / 128
  16.1.4 The Marseilles Group / 128
  16.1.5 The San Diego Group / 128
  16.1.6 The Helsinki Group / 129
16.2 Summary / 129
Further Reading / 129

17. Ovarian Cancer

17.1 Microarray Studies of Ovarian Cancer / 130
  17.1.1 Duke University Medical Center / 130
  17.1.2 The Stanford Group / 131
  17.1.3 The London Group / 131
  17.1.4 The UCLA Group / 131
  17.1.5 The NIH Group / 131
  17.1.6 The Novartis Group / 132
17.2 Summary / 132
Further Reading / 132
18. Prostate Cancer

18.1 Microarray Studies of Prostate Cancer / 134
   18.1.1 University of Michigan Medical School / 134
   18.1.2 Memorial Sloan-Kettering Cancer Center / 134
   18.1.3 The NIH Group / 135
   18.1.4 The Harvard Group / 135
   18.1.5 The Australian Group / 135
   18.1.6 The Stanford Group / 136

18.2 Summary / 136
   Further Reading / 136

19. Melanoma

19.1 Microarray Studies of Melanoma / 138
   19.1.1 Memorial Sloan-Kettering Cancer Center / 138
   19.1.2 The NIH Group / 138

19.2 Summary / 139
   Further Reading / 139

20. Brain Tumors

20.1 Microarray Studies of Brain Tumors / 144
   20.1.1 Harvard Medical School / 144
   20.1.2 M. D. Anderson Cancer Center / 144
   20.1.3 UCLA School of Medicine / 144
   20.1.4 The Lausanne Group / 144
   20.1.5 Deutsches Krebsforschungszentrum / 144
   20.1.6 Children’s Hospital, Boston / 145

20.2 Summary / 145
   Further Reading / 145

21. Organ or Tissue Specific Classification

   Further Reading / 148

22. Sample Collection and Stability

   22.1 Tissue Samples / 149
      22.1.1 Stability of Tissue After Surgical Removal / 149
      22.1.2 Stability of Sample in Storage / 150
      22.1.3 Paraffin-Embedded Tissue Samples / 150
   22.2 Blood Samples / 150
   22.3 Sample Heterogeneity / 151
## CONTENTS

22.4 Summary / 151
Further Reading / 151

<table>
<thead>
<tr>
<th>References</th>
<th>152</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index</td>
<td>183</td>
</tr>
</tbody>
</table>
Preface

A new technology is about to enter cancer diagnostics. DNA microarrays are currently showing great promise in all the medical research projects to which they are being applied. This book presents the current status of the area as well as reviews and summaries of the results from specific cancer types. For types where several comparable studies have been published, a meta-analysis of the results is presented.

This book is intended for a wide audience from the practicing physician to the statistician. Both will find the chapters where I review areas of their expertise superficial, but I hope they will find other introductory chapters useful in understanding the many aspects of microarray technology applied to cancer diagnostics.

I first describe the current state of the technology as well as emerging technologies. Then I describe the statistical analysis that is necessary to interpret the data. Next I cover some of the major human cancer types where microarrays have been applied with success, including studies that I have been a part of. I conclude that for several cancer types the results are so good and consistent that DNA microarrays are ready to be deployed in clinical practice. The clinical application in question is helping to select patients for adjuvant chemotherapy after surgery by determining the prognosis more accurately than what is possible today.

Chapters 1–6 on technology and statistical analysis and Chapters 9 and 10 on chip design and software are updated and expanded versions of chapters in my previous Wiley book, Guide to Analysis of DNA Microarray Data, Second Edition (2004). The remaining 13 chapters are new.

Steen Knudsen

Birkerød, Denmark
October 2005
Acknowledgments

My previous books, on which some of the chapters in this book are based, were written while I worked for the Technical University of Denmark. I am grateful to the University leadership, the funding agencies, my group members, and collaborators for their significant role in that endeavor. The review of individual cancer types, to a large extent based on information from the American Cancer Society, was also written during my employment at the Technical University of Denmark.

The remaining chapters were written while I was employed by the Medical Prognosis Institute. The original research on meta-analysis of cancer classification as well as subnetwork mapping of individual cancer types has been patented by Medical Prognosis Institute, and I am grateful for being allowed to present the results here.

S. K.
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) (Figure 1.1). 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 radiolabeled 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 radiolabeled probe present in the band, which again depends on the amount of messenger. So this method allows 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 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. I doubt there is any disease that does not in some way affect the expression of genes in some cells. That is the basis for this book. By applying DNA microarrays to human tissue or human cells, we can learn about disease and characterize disease at a much more detailed level than what was previously possible. Cancer has turned out to be the disease that has attracted the most focus. One reason is that it is possible to obtain a tissue sample from the tumor during surgery. This tissue sample is then used to measure gene expression with a DNA microarray.

DNA arrays can also be used to hybridize to DNA. In that case they can distinguish between different alleles (mutations) in DNA, some of which affect cancer. DNA chips have been developed for detecting mutations in the human TP53 tumor suppressor gene (Ahrendt et al., 1999; Wikman et al., 2000; Spicker et al., 2002). Likewise, a DNA chip for detecting mutations in the human CYP450 genes, important for metabolizing common drugs, has been developed. These genotyping or resequencing chips will not be covered further in this book, which focuses exclusively on gene expression in cancer.

More recently, arrays have been used for comparative genomic hybridization (CGH) to reveal deletion or duplication of chromosomal regions in cancer (Pollack et al., 2002; Douglas et al., 2004; Blaveri et al., 2005a; Jones et al., 2005). While the results have been very promising, they are limited in number and will not be covered in this book. Similarly, arrays have recently been used to detect small noncoding RNAs (miRNA) and a correlation to cancer has been found (Lu et al., 2005). This will not be covered further in this book.