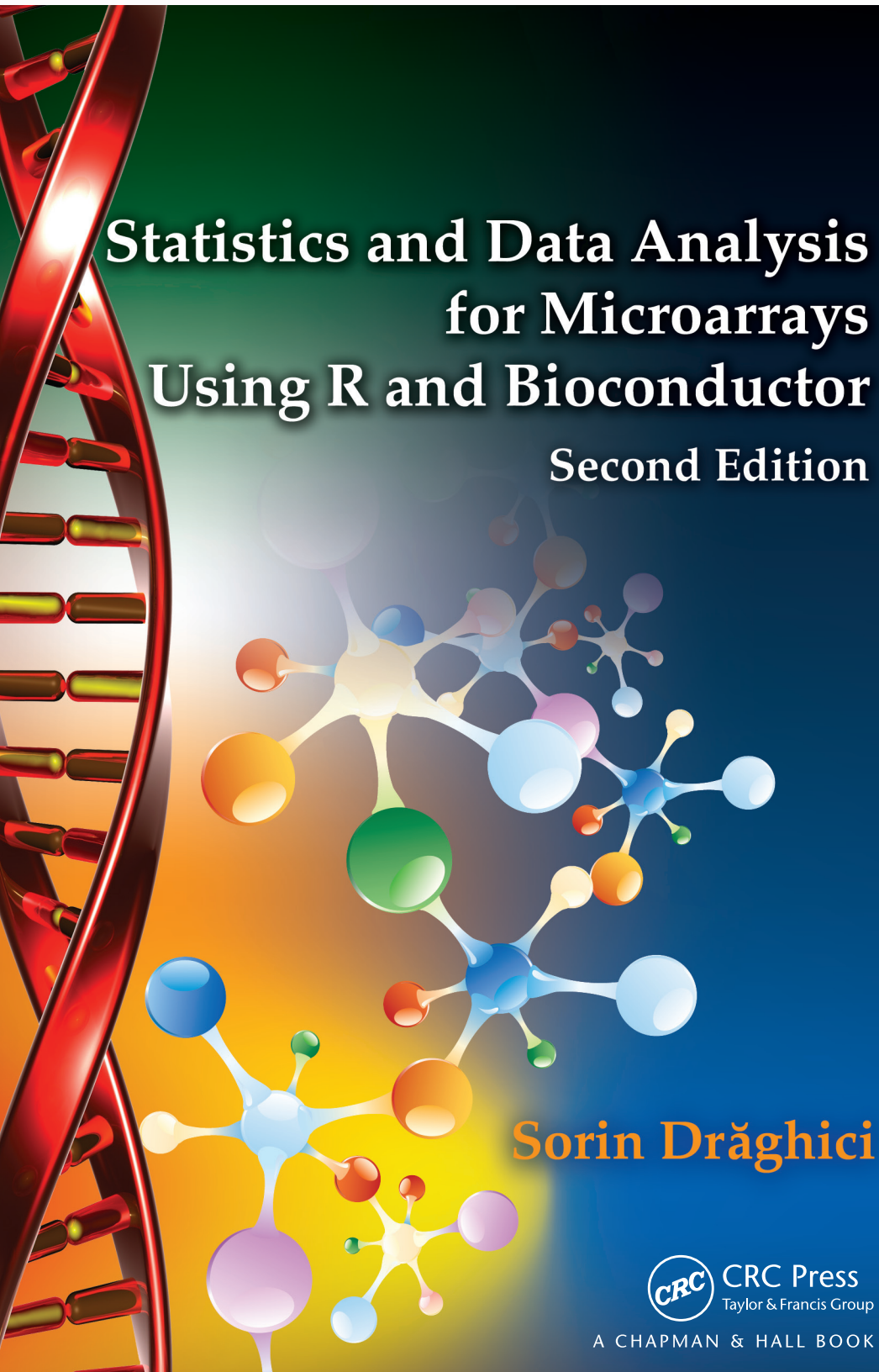
Chapman & Hall/CRC
Mathematical and Computational Biology Series



Statistics and Data Analysis for Microarrays Using R and Bioconductor

Second Edition

Sorin Drăghici

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Statistics and Data Analysis for Microarrays Using R and Bioconductor

Second Edition

CHAPMAN & HALL/CRC
Mathematical and Computational Biology Series

Aims and scope:
This series aims to capture new developments and summarize what is known over the entire spectrum of mathematical and computational biology and medicine. It seeks to encourage the integration of mathematical, statistical, and computational methods into biology by publishing a broad range of textbooks, reference works, and handbooks. The titles included in the series are meant to appeal to students, researchers, and professionals in the mathematical, statistical and computational sciences, fundamental biology and bioengineering, as well as interdisciplinary researchers involved in the field. The inclusion of concrete examples and applications, and programming techniques and examples, is highly encouraged.

Series Editors

Proposals for the series should be submitted to one of the series editors above or directly to:
CRC Press, Taylor & Francis Group
4th, Floor, Albert House
1-4 Singer Street
London EC2A 4BQ
UK

Published Titles

Algorithms in Bioinformatics: A Practical Introduction
Wing-Kin Sung

Bioinformatics: A Practical Approach
Shui Qing Ye

Biological Computation
Ehud Lamm and Ron Unger

Biological Sequence Analysis Using the SeqAn C++ Library
Andreas Gogol-Döring and Knut Reinert

Cancer Modelling and Simulation
Luigi Preziosi

Cancer Systems Biology
Edwin Wang

Cell Mechanics: From Single Scale-Based Models to Multiscale Modeling
Arnaud Chauvière, Luigi Preziosi, and Claude Verdier

Clustering in Bioinformatics and Drug Discovery
John D. MacCuish and Norah E. MacCuish

Combinatorial Pattern Matching Algorithms in Computational Biology Using Perl and R
Gabriel Valiente

Computational Biology: A Statistical Mechanics Perspective
Ralf Blossey

Computational Hydrodynamics of Capsules and Biological Cells
C. Pozrikidis

Computational Neuroscience: A Comprehensive Approach
Jianfeng Feng

Data Analysis Tools for DNA Microarrays
Sorin Draghici

Differential Equations and Mathematical Biology, Second Edition
D.S. Jones, M.J. Plank, and B.D. Sleeman

Dynamics of Biological Systems
Michael Small

Engineering Genetic Circuits
Chris J. Myers

Exactly Solvable Models of Biological Invasion
Sergei V. Petrovskii and Bai-Lian Li

Gene Expression Studies Using Affymetrix Microarrays
Hinrich Göhlmann and Willem Talloen

Glycome Informatics: Methods and Applications
Kiyoko F. Aoki-Kinoshita

Handbook of Hidden Markov Models in Bioinformatics
Martin Gollery

Introduction to Bioinformatics
Anna Tramontano

Introduction to Bio-Ontologies
Peter N. Robinson and Sebastian Bauer

Introduction to Computational Proteomics
Golan Yona

Introduction to Proteins: Structure, Function, and Motion
Amit Kessel and Nir Ben-Tal

An Introduction to Systems Biology: Design Principles of Biological Circuits
Uri Alon

Kinetic Modelling in Systems Biology
Oleg Demin and Igor Goryanin

Knowledge Discovery in Proteomics
Igor Jurisica and Dennis Wigle

Meta-analysis and Combining Information in Genetics and Genomics
Rudy Guerra and Darlene R. Goldstein

Methods in Medical Informatics: Fundamentals of Healthcare Programming in Perl, Python, and Ruby
Jules J. Berman

Modeling and Simulation of Capsules and Biological Cells
C. Pozrikidis

Niche Modeling: Predictions from Statistical Distributions
David Stockwell

Published Titles (continued)

Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems
Qiang Cui and Ivet Bahar

Optimal Control Applied to Biological Models
Suzanne Lenhart and John T. Workman

Pattern Discovery in Bioinformatics: Theory & Algorithms
Laxmi Parida

Python for Bioinformatics
Sebastian Bassi

Spatial Ecology
Stephen Cantrell, Chris Cosner, and Shigui Ruan

Spatiotemporal Patterns in Ecology and Epidemiology: Theory, Models, and Simulation
Horst Malchow, Sergei V. Petrovskii, and Ezio Venturino

Statistics and Data Analysis for Microarrays Using R and Bioconductor, Second Edition
Sorin Drăghici

Stochastic Modelling for Systems Biology
Darren J. Wilkinson

Structural Bioinformatics: An Algorithmic Approach
Forbes J. Burkowski

The Ten Most Wanted Solutions in Protein Bioinformatics
Anna Tramontano

Chapman & Hall/CRC Mathematical and Computational Biology Series

Statistics and Data Analysis for Microarrays Using R and Bioconductor

Second Edition

Sorin Drăghici

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 2011909

International Standard Book Number-13: 978-1-4398-0976-1 (eBook - PDF)

To Jeannette, my better half,
to Tavi, who brightens every day of my life,
and to Althea, whom I miss every day we are not together

This page intentionally left blank

Contents

This page intentionally left blank

List of Figures

List of Tables

Art is science made clear.

—Jean Cocteau

Any good poet, in our age at least, must begin with the scientific view of the world; and any scientist worth listening to must be something of a poet, must possess the ability to communicate to the rest of us his sense of love and wonder at what his work discovers.

—Edward Abbey, The Journey Home

The most erroneous stories are those we think we know best - and therefore never scrutinize or question.

—Stephen Jay Gould

My definition of an expert in any field is a person who knows enough about what's really going on to be scared.

—P.J. Plauger

This page intentionally left blank

Preface

Although the industry once suffered from a lack of qualified targets and candidate drugs, lead scientists must now decide where to start amidst the overload of biological data. In our opinion, this phenomenon has shifted the bottleneck in drug discovery from data collection to data analysis, interpretation and integration.

—Life Science Informatics, UBS Warburg Market Report, 2001

One of the most promising tools available today to researchers in life sciences is the microarray technology. Typically, one DNA array will provide hundreds or thousands of gene expression values. However, the immense potential of this technology can only be realized if many such experiments are done. In order to understand the biological phenomena, expression levels need to be compared between species or between healthy and ill individuals or at different time points for the same individual or population of individuals. This approach is currently generating an immense quantity of data. Buried under this humongous pile of numbers lays invaluable biological information. The keys to understanding phenomena from fetal development to cancer may be found in these numbers. Clearly, powerful analysis techniques and algorithms are essential tools in mining these data. However, the computer scientist or statistician that does have the expertise to use advanced analysis techniques usually lacks the biological knowledge necessary to understand even the simplest biological phenomena. At the same time, the scientist having the right background to formulate and test biological hypotheses may feel a little uncomfortable when it comes to analyzing the data thus generated. This is because the data analysis task often requires a good understanding of a number of different algorithms and techniques and most people usually associate such an understanding with a background in mathematics, computer science, or statistics.

Because of the huge amount of interest around the microarray technology, there are quite a few books available on this topic. Many of the few available texts concentrate more on the wet lab techniques than on the data analysis aspects. There are several books that review the topic in a somewhat superficial manner, covering everything there is to know about data analysis of microarrays in a couple of hundred pages or less. Other available books focus on excruciating details that only developers of analysis packages find useful. Others are simple proceedings of conferences, gathering together unrelated

papers that focus on very specific aspects and topics. Overall, I felt there was a need for a good, middle-of-the-road book that would cover topics in sufficient details to make it possible for readers to really understand what is going on, but without overwhelming details or intimidating heavy formalisms or notations.

At the same time, the R environment has started to dominate heavily everything that is done in terms of data analysis in this area. Again, while many good books on R as a programming and analysis language are available, I felt that a book that would allow the reader to become competent in the analysis of microarray data by providing: i) everything needed to learn the basics of R, ii) the basics of the microarray technology, as well as iii) the understanding necessary in order to apply the right tools to the right problems would be beneficial.

Audience and prerequisites

The goal of this book is to fulfill this need by presenting the main computational techniques available in a way that is useful to both life scientists and analytical scientists. The book tries to demolish the imaginary concrete wall that separates biology and medicine from computer science and statistics and allow the biologist to be a refined user of the available techniques, as well as be able to communicate effectively with computer scientists and statisticians designing new analysis techniques. The intended audience includes as a central figure the researcher or practitioner with a background in the life sciences that needs to use computational tools in order to analyze data. At the same time, the book is intended for the computer scientists or statisticians who would like to use their background in order to solve problems from biology and medicine. The book explains the nature of the specific challenges that such problems pose as well as various adaptations that classical algorithms need to undergo in order to provide good results in this particular field.

Finally, it is anticipated that there will be a shift from the classical compartmented education to a highly interdisciplinary approach that will form people with skills across a range of disciplines crossing the borders between traditionally unrelated fields, such as medicine or biology and statistics or computer science. This book can be used as a textbook for a senior undergraduate or graduate course in such an interdisciplinary curriculum. The book is suitable for a data analysis and data mining course for students with a background in biology, molecular biology, chemistry, genetics, computer science, statistics, mathematics, etc.

Useful prerequisites for a biologist include elementary calculus and algebra. However, the material is designed to be useful even for readers with a shaky mathematical foundation since those elements that are crucial for the topic are fully discussed. Useful prerequisites for a computer scientist or mathematician include some elements of genetics and molecular biology. Once

again, such knowledge is useful but not required since the essential aspects of the technology are covered in the book.

Aims and contents

The first and foremost aim of this book is **to provide a clear and rigorous description of the algorithms without overwhelming the reader with the usual cryptic notation or with too much mathematical detail**. The presentation level is appropriate for a scientist with a background in life sciences. Little or no mathematical training is needed in order to understand the material presented here. Those few mathematical and statistical facts that are really needed in order to understand the techniques are completely explained in the book at a level that is fully accessible to the non-mathematically minded reader. The goal here was to keep the level as accessible as possible. The mathematical apparatus was voluntarily limited to the very basics. The most complicated mathematical symbol throughout the book is the sum of n terms: $\sum_{i=1}^{n} x_i$. In order to do this, certain compromises had to be made. The definitions of many statistical concepts are not as comprehensive as they could be. In certain places, giving the user a powerful intuition and a good understanding of the concept took precedence over the exact, but more difficult to understand, formalism. This was also done for the molecular biology aspects. Certain cellular phenomena have been presented in a simplified version, leaving out many complex phenomena that we considered not to be absolutely necessary in order to understand the big picture.

A second specific aim of the book is **to allow a reader to learn the R environment and programming language using a hands-on and example-rich approach**. From this perspective, the book should be equally useful to a large variety of readers with very different backgrounds. No previous programming experience is required or expected from the reader. The book includes chapters that describe everything from the basic R commands and syntax to rather sophisticated procedures for quality control, normalization, data analysis and machine learning. Everything from the simplest commands to the most complex procedures is illustrated with R code. All analysis results shown in the book are actual results produced by the code shown in the text and therefore, all code shown is free from spelling or syntax errors.

A third specific aim of the book is **to allow a microarray user to be in a position to make an informed choice as to what data analysis technique to use in a given situation, even if using other analysis packages**. The existing software packages usually include a very large number of techniques, which in turn use an even larger number of parameters. Thus, the biologist trying to analyze DNA microarray data is confronted with an overwhelming number of possibilities. Such flexibility is absolutely crucial because each data set is different and has specific particularities that must be taken into account when selecting algorithms. For example, data sets obtained in different laboratories have different characteristics, so the choice of normal-

ization procedures is very important. However, such wealth of choices can be overwhelming for the life scientist who, in most cases, is not very familiar with all intricacies of data analysis and ends up by always using the default choices. This book is designed to help such a scientist by emphasizing at a high level of abstraction the characteristics of various techniques in a biological context.

As a text designed to bridge the gap between several disciplines, the book includes chapters that would give all the necessary information to readers with a variety of backgrounds. The book is divided into two parts. The first part is designed to offer an overview of microarrays and to create a solid foundation by presenting the elements from statistics that constitute the building blocks of any data analysis. The second part introduces the reader to the details of the techniques most commonly used in the analysis of microarray data.

Chapter 2 presents a short primer on the central dogma of molecular biology and why microarrays are useful. This chapter is aimed mostly at analytical scientists with no background in life sciences. **Chapter 3** briefly presents the microarray technology. For the computer scientist or statistician, this constitutes a microarray primer. For the microarray user, this will offer a bird's-eye view perspective on several techniques emphasizing common as well as technology-specific issues related to data analysis. This is useful since many times the users of a specific technology are so engulfed in the minute details of that technology that they might not see the forest for the trees.

Chapter 4 discusses a number of important issues related to the reliability and reproducibility of microarray data. This discussion is important both for the life scientist who needs to understand the limitations of the technology used, as well as for the computer scientist or statistician who needs to understand the intrinsic level of noise present in this type of data.

Chapter 5 constitutes a short primer on digital imaging and image processing. This chapter is mostly aimed at the life scientists or statisticians who are not familiar with digital image processing.

Chapter 6 is an introduction to the R programming language. This chapter discusses basic concepts from the installation of the R environment to the basic syntax and concepts of R.

Chapter 7 presents the Bioconductor project and briefly illustrates its capabilities with some simple examples. This chapter was contributed by one of the founders of the Bioconductor project, Vincent Carey, currently an Associate Professor of Medicine (Biostatistics) at Harvard Medical School, and an Associate Biostatistician in the Department of Medicine at the Brigham and Women's Hospital in Boston.

Chapters 8, **9**, and **11** focus on some elementary statistics notions. These chapters will provide the biologist with a general perspective on issues very intimately related to microarrays. The purpose here is to give only as much information as needed in order to be able to make an informed choice during the subsequent data analysis. The aim of the discussion here is to put things in the perspective of somebody who analyzes microarray data rather than offer a full treatment of the respective statistical notions and techniques. **Chapter 9**

discusses several important distributions, **Chapter 11** discusses the classical hypothesis testing approach, and **Chapter 12** applies it to microarray data analysis. **Chapter 10** uses R to illustrated the basic statistical tools available in R for descriptive statistics and basic built-in distributions.

Chapter 13 presents the family of ANalysis Of VAriance methods intensively used by many researchers to analyze microarray data. **Chapter 14** discusses the more general linear models and illustrates them in R. **Chapter 15** uses some of the ANOVA and linear model approaches in the discussion of various techniques for experiment design.

Chapter 16 discusses several issues related to the fact that microarrays interrogate a very large number of genes simultaneously and its consequences regarding data analysis.

Chapters 17 and **18** present the most widely used tools for microarray data analysis. In most cases, the techniques are presented using real data. Chapter 17 includes several techniques used in exploratory analysis, when there is no known information about the problem and the task is to identify relevant phenomena as well as the parameters (genes) that control them. The main techniques discussed here include box plots, histograms, scatter plots, volcano plots, time series, principal component analysis (PCA), and independent component analysis (ICA). The clustering techniques described in Chapter 18 include K-means, hierarchical clustering, biclustering, partitioning-around-medoids, and self-organizing feature maps. Again, the purpose here is to explain the techniques in an unsophisticated yet rigorous manner. The all-important issue of when to use a specific technique is discussed on various examples emphasizing the strengths and weaknesses of each individual technique.

Chapter 19 discusses specific quality control issues characteristic to Affymetrix and Illumina data. These are illustrated using R functions and packages applied on real data sets. Tools such as intensity distributions, box plots, RNA degradation curves, and quality control metrics are used to illustrate problems ranging from array saturation, to RNA degradation, and annotation issues. Plots illustrating various problems are shown side-by-side with plots showing clean data such that the reader can understand and learn what to look for in such plots.

Chapter 20 concentrates on data preparation issues. Although such issues are crucial for the final results of the data mining process, they are often ignored. Issues such as color swapping, color normalization, background correction, thresholding, mean normalization, etc., are discussed in detail. This chapter will be extremely useful both to the biologist, who will become aware of the different numerical aspects of the various preprocessing techniques, and to the computer scientist, who will gain a deeper understanding of various biological aspects, motivations, and meanings behind such preprocessing. Again, all normalization issues are illustrated using R functions and packages applied on real data sets.

Chapter 21 presents several methods used to select differentially regulated genes in comparative experiments.

Chapter 22 discusses the Gene Ontology, including its goal, structure, annotations, and some statistics about the data currently available in it. **Chapter 23** shows how GO can be used to translate lists of differentially expressed genes into a better understanding of the underlying biological phenomena. Just when you think this is easy, **Chapter 24** comes to tell you about the many mistakes and issues that could appear in this GO profiling. **Chapter 25** reviews more than a dozen tools that are currently available for this type of functional analysis.

Chapter 26 somehow reverses the direction considering the problem of how to select the microarrays that are best suited for investigating a given biological hypothesis.

Chapter 27 discusses some of the problems that can be caused by the fact that the same biological entity may have different IDs in different public databases. Some tools that allow a mapping from one type of ID to another are discussed and compared.

Chapter 28 takes the analysis to the next level, using a systems biology approach that aims to take into consideration the way genes are known to interact with each other. This type of knowledge is captured in collections of signaling pathways available from various sources. This chapter discusses various approaches currently available for the analysis of signaling pathways.

Chapter 29 is a brief review of several machine learning techniques that are widely used with microarray data. Since unsupervised methods are discussed in **Chapter 18**, this chapter focuses on supervised methods including linear discriminants, feed-forward neural networks, and support vector machines.

Finally, the last chapter of the book presents some conclusions as well as a brief presentation of some novel techniques expected to have a great impact on this field in the near future.

Road map

This book can be used in several ways, depending on the background of the reader and the goals pursued. The chapters can be combined in various ways, allowing an instructor to tailor a course to the specific background and expectations of a given audience.

Some of the courses (with or without a laboratory component) that can be easily taught using this book include:

1. Introduction to statistics: Chapters 8, 9, 11, 12, 13, 14, 15, 16
2. Introduction to R and Bioconductor: Chapters 6, 7, 10, 14, 17, 18, 29

3. Microarray data analysis for life scientists: Chapters 3 – 30
4. Microarray data analysis for computer scientists (including R and Bioconductor): Chapter 2 – 30
5. Quality control and normalization techniques: Chapters 19, 20
6. Interpretation of high-throughput data: GO profiling and pathway analysis: Chapters 22 – 28

This book focuses on R and Bioconductor. The reader is advised to install the software and actually use it to perform the analysis steps discussed in the book. The accompanying CD includes all code used throughout the book.

Acknowledgments

The author of this book has been supported by the Biological Databases Program of the National Science Foundation under grants number NSF-0965741 and NSF-0234806, the Bioinformatics Cell, MRMC US Army – DAMD17-03-2-0035, National Institutes of Health – grant numbers 1R01DK089167-01, R01-NS045207-01 and R21-EB000990-01, and Michigan Life Sciences Corridor grant number MLSC-27. Many thanks to Sylvia Spengler, Director of the Biological Databases program, National Science Foundation, Salvatore Sechi, Program Manager NIDDKD, Peter McCartney, Program Director, Division of Biological Infrastructure, National Science Foundation, Peter Lyster, Program Director in the Center for Bioinformatics and Computational Biology, NIGMS without whose support our research and this book would not have been possible.

I would like to express my gratitude to Dr. Robert Romero, Chief of the Perinatology Research Branch of the National Institute of Child Health and Human Development, who has been a tremendous source of inspiration and a role model in my research. I am also very grateful for his strong support which allowed me to dedicate more time to my research and to other scholarly endeavors such as writing this book.

My deepest gratitude goes to Robert J. Sokol, M.D., The John M. Malone, Jr., M.D., Endowed Chair, and Director of the C.S. Mott Center for Human Growth and Development, Distinguished Professor of Obstetrics and Gynecology in the Wayne State University School of Medicine. His efforts and support were the main reasons that kept me at Wayne State when I had great opportunities elsewhere. Also, his generosity in creating the endowed chair in systems biology that bears his name – and that I am currently holding – made possible the creation of the strong research group of which I am a part.

My colleague Adi Laurentiu Tarca had a substantial contribution to this

book. This consisted of a first draft of the linear models chapter, as well as numerous code snippets used throughout several other chapters to illustrate various concepts and techniques. He also contributed with a very thorough critical reading of many other chapters followed by very useful comments and suggestions.

My thanks also go to Vincent Carey, Associate Professor of Medicine (Biostatistics) at Harvard Medical School, and an Associate Biostatistician in the Department of Medicine at the Brigham and Women's Hospital in Boston. Vincent is one of the founders of the Bioconductor project and contributed Chapter 7 to this book. He was also very helpful in answering several questions regarding Bioconductor and various packages used throughout this book.

My Ph.D. students Calin Voichita, Michele Donato, Cristina Mitrea, Dorina Twigg, Munir Islam, Rebecca Tagett, and my visiting postdoctoral researcher Josep Maria Mercader (currently back at his home institution – Barcelona Supercomputer Center) contributed enormously to this book, first by accepting to be my guinea pigs for the examples and explanations used in the book, and then by proof-reading all 1,081 pages of this book – several times. In spite of their tremendous efforts, typos and small errors probably continue to exist here and there. If this is the case, the fault is entirely mine. I am sure that for each such surviving typo, somewhere in my inbox there is an email from one of them, bringing it to my attention.

My research associate, Zhonghui Xu, had a substantial contribution to this book by writing various R routines and snippets of code, in particular most of those in the chapters on normalization and quality control.

Gary Chase kindly reviewed the chapters on statistics (for the first incarnation of this book) and did so on incredibly short deadlines that allowed the book to be published on time. His comments were absolutely invaluable and made the book stronger than I initially conceived it.

I owe a lot to my colleague, friend and mentor, Michael A. Tainsky. Michael taught me everything I know about molecular biology and I continue to be amazed by his profound grasp of some difficult data analysis concepts. I have learned a lot during our collaboration over the past 10 years. Michael also introduced me to Judy Abrams who provided a lot of encouragement and has been a wonderful source of fresh and exciting ideas, many of which ended up in grant proposals. My colleague and friend Steve Krawetz was the first one to identify the need for something like Onto-Express (Chapter 23). During my leave of absence, he continued to work with my students and participated actively in the creation of our first Onto-Express prototype. Our collaboration over the past few years produced a number of great ideas and a few papers. Thanks also go to Jeff Loeb for our productive interaction as well as for our numerous discussions on corrections for multiple experiments.

Many thanks to Otto Muzik for sticking to the belief that our collaboration is worthwhile through the first, less than productive, year and for being such a wonderful colleague and friend. In spite of some personal differences, Otto did not hesitate to offer his help when my father was diagnosed with cancer.

I am very grateful to him for his offer to help, even though my father passed away before he could take advantage of Otto's kind offer.

Jim Granneman and Bob Mackenzie knocked one day on my door looking for some help analyzing their microarray data. One year later, we were awarded a grant of over 3.5 million from the Michigan Life Science Corridor. This allowed me to shift some of my effort to this book. Currently, Jim and Bob are my coPIs on a 1.5 million NIH grant which is funding some very exciting research. Thank you all for involving me in your work.

I am grateful to Soheil Shams, Bruce Hoff, Anton Petrov and everybody else with whom I interacted during my one year sabbatical at BioDiscovery. Bruce, Anton, and I worked together at the development of the ANOVA based noise sampling method described in Chapter 21 based on some initial experiments by Xiaoman Li. Figures 18.6, 18.12, and 18.15 were initially drawn by Bruce.

During my stay in Los Angeles, I met some top scientists, including Michael Waterman and Wing Wong. I learned a lot from our discussions and I am very grateful for that. Mike Waterman provided very useful comments and suggestions on the noise sampling method in Chapter 21 and later, on OntoExpress described in Chapter 23. Wing Wong and his students at UCLA and Harvard have done a great job with their software dChip, which is a great first stop every time one has to normalize Affymetrix data.

Gary Churchill provided useful comments on the work related to the ANOVA based noise sampling technique presented in Chapter 21. The discussions with him and Kathy Kerr helped us a lot. Also, their ANOVA software for microarray data analysis represent a great tool for the research community. We used some of their software to generate the images illustrating the LOWESS transform in Chapter 20.

John Quackenbush kindly provided a wonderful sample data set that we used in the PCA examples in Chapter 17. He and his colleagues at TIGR designed, implemented and made available to the community a series of very nice tools for microarray data analysis including the Multiple Experiment Viewer (MEV) used to generate some images for Chapter 17.

About the author

Sorin Drăghici has obtained his B.Sc. and M.Sc. degrees in Computer Engineering from "Politehnica" University in Bucharest, Romania followed by a Ph.D. degree in Computer Science from University of St. Andrews, United Kingdom (third oldest university in UK after Oxford and Cambridge). Besides this book, he has published two other books, several book chapters, and over 100 peer-reviewed journal and conference papers. He is inventor or co-inventor on several patent applications related to biotechnology, high-throughput meth-

ods and systems biology. He is currently an editor of IEEE/ACM Transactions on Computational Biology and Bioinformatics, the Journal of Biomedicine and Biotechnology, and the International Journal of Functional Informatics and Personalized Medicine. He is also active as a journal reviewer for 15 international technical journals related to bioinformatics and computational biology, as well as a regular NSF and NIH panelist on biotechnology and bioinformatics topics.

Currently, Sorin Drăghici is holding the Robert J. Sokol, MD Endowed Chair in Systems Biology in the Department of Obstetrics and Gynecology in the School of Medicine, as well as joint appointments as Professor in the Department of Clinical and Translational Science, School of Medicine, and the Department of Computer Science, College of Engineering, Wayne State University. He is also the Chief of the Bioinformatics and Data Analysis Section, Perinatology Research Branch, National Institute for Child Health and Development, NIH, as well as the head of the Intelligent Systems and Bioinformatics Laboratory (ISBL, http://vortex.cs.wayne.edu).

Disclaimer

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation, National Institutes of Health, any other funding agency, nor those of any of the people mentioned above.

Chapter 1

Introduction

If we begin with certainties, we shall end in doubts; but if we begin with doubts, and are patient in them, we shall end in certainties.

—*Francis Bacon*

1.1 Bioinformatics – an emerging discipline

Life sciences are currently at the center of an informational revolution. Dramatic changes are being registered as a consequence of the development of techniques and tools that allow the collection of biological information at an unprecedented level of detail and in extremely large quantities. The human genome project is a compelling example. Initially, the plan to sequence the human genome was considered extremely ambitious, on the border of feasibility. The first serious effort was planned over 15 years at a cost of $3 billion. Soon after, the schedule was revised to last only 5 years. Eventually, the genome was sequenced in less than 3 years, at a cost much lower than initially expected [361]. The nature and amount of information now available open directions

of research that were once in the realm of science fiction. Pharmacogenomics [359], molecular diagnostics [86, 125, 260, 360, 376, 455] and drug target identification [302] are just a few of the many areas [163] that have the potential to use this information to change dramatically the scientific landscape in the life sciences.

During this informational revolution, the data-gathering capabilities have greatly surpassed the data analysis techniques. If we were to imagine the Holy Grail of life sciences, we might envision a technology that would allow us to fully understand the data at the speed at which it is collected. Sequencing, localization of new genes, functional assignment, pathway elucidation, and understanding the regulatory mechanisms of the cell and organism should be seamless. Ideally, we would like knowledge manipulation to become tomorrow the way goods manufacturing is today: high automatization producing more goods, of higher quality, and in a more cost-effective manner than manual production. In a sense, knowledge manipulation is now reaching its pre-industrial age. Our farms of sequencing machines and legions of robotic arrayers can now produce massive amounts of *data* but using them to manufacture highly processed pieces of *knowledge* still requires skilled masters painstakingly forging through small pieces of raw data one at a time. The ultimate goal in life science research is to automate this knowledge discovery process.

Bioinformatics is the field of research that presents both the opportunity and the challenge to bring us closer to this goal. Bioinformatics is an emerging discipline situated at the interface between analytical sciences such as statistics, mathematics, and computer science on one side, and biological sciences such as molecular biology, genomics, and proteomics on the other side. Initially, the term bioinformatics was used to denote very specific tasks such as the activities related to the storage of data of biological nature in databases. As the field evolved, the term has started to also encompass algorithms and techniques used in the context of biological problems. Although there is no universally accepted definition of bioinformatics, currently the term denotes a field concerned with the application of information technology techniques, algorithms, and tools to solve problems in biological sciences. The techniques currently used have their origins in a number of areas, such as computer science, statistics, mathematics, etc. Essentially, **bioinformatics** is the science of refining biological information into biological knowledge using computers. Sequence analysis, protein structure prediction, and dynamic modeling of complex biosystems are just a few examples of problems that fall under the general umbrella of bioinformatics. However, new types of data have started to emerge. Examples include protein–protein interactions, protein–DNA interactions, signaling and biochemical pathways, population-scale sequence data, large-scale gene expression data, and ecological and environmental data [191].

A subfield of particular interest today is genomics. The field of **genomics** encompasses investigations into the structure and function of very large number of genes undertaken in a simultaneous fashion. The term genomics comes from **genome**, which is the entire set of genes in a given organism. **Structural**

genomics includes the genetic mapping, physical mapping, and sequencing of genes for entire organisms (genomes). **Comparative genomics** deals with extending the information gained from the study of some organisms to other organisms. **Functional genomics** is concerned with the role that individual genes or subsets of genes play in the development and life of organisms.

Just as genomics refers to the large-scale studies involving the properties and functions of many genes, **proteomics** is concerned with the large-scale study of proteins. The **proteome** of an organism is the set of all proteins in the given organism. Following the same lines of linguistic synthesis, the "-omics" suffix has been appended to a number of other terms generating a large variety of novel terms. Among those, some have been well accepted by the community, such as transcripts → **transcriptome** → **transcriptomics**, metabolite → **metabolome** → **metabolomics**, etc. In all cases, the "X-ome," represents the set of all entities of type X in a given organism, while the "X-omics" represents the field of research studying them using high-throughput approaches.

Currently, our understanding of the role played by various genes and their interactions seems to be lagging far behind the knowledge of their sequence information. Table 1.1 presents some data that reflect the relationship between sequencing an organism and understanding the role of its various genes [311]. The yeast is an illustrative example. Although the 6,200 genes of its genome have been known since 1997, only approximately 94% of them have inferred functions. The situation is similar for *E. coli*, *C. elegans*, *Drosophila*, and *Arabidopsis*.[1]

Most researchers agree that the challenge of the near future is to analyze, interpret, and understand all data that are being produced [56, 148, 290, 440]. In essence, the challenge faced by the biological scientists is to use the large-scale data that are being gathered to discover and understand fundamental biological phenomena. At the same time, the challenge faced by computer scientists is to develop new algorithms and techniques to support such discoveries [191].

The explosive growth of computational biology and bioinformatics has just started. Biotechnology and pharmaceutical companies are channeling many resources towards bioinformatics by starting informatics groups. The tendency has been noted by the academic world, and there are a number of universities that have declared bioinformatics a major research priority. The need for bioinformatics-savvy people as well as experts in bioinformatics is enormous and will continue to accentuate in the near future. Two important phenomena are at play here. On the one hand, modern life science research has irreversibly adopted the use of very large throughput technologies such as DNA microarrays, mass-spectrometry, high-throughput sequencing, etc. These techniques generate terabytes of data on a daily basis. On the other hand, the academic

[1] Even genomes that are considered substantially complete, in reality may still have small gaps [6]. Until we understand better the function of various genes, we cannot discount the functional relevance of the genetic material in those gaps.

Organism	Number of genes	Genes with inferred function	Genome Completion date
S. cerevisiae	6,201	94%	1996 [181]
E. coli	4,467	62%	1997 [55]
C. elegans	21,185	63%	1998 [448]
D. melanogaster	18,462	82%	1999 [6]
A. thaliana	33,264	76%	2000 [370]
Homo sapiens	39,920	46%	2001 [437]
B. anthracis	5,415	80%	2002 [351]
Rattus norvegicus	37,533	37%	2004 [173]

TABLE 1.1: The unbalance between obtaining the data and understanding it. Although the complete genomes of several simpler organisms are available, understanding the role of various genes lags far behind. *Saccharomyces cerevisiae* is the baker's yeast; *Escherichia coli* is a bacterium that lives in the gut of worm-blooded animals (and some of its varieties sometimes infect the human food chain); *C. elegans* is a nematode (worm); *Drosophila melanogaster* is the fruit fly; *Arabidopsis thaliana* is a plant; *Bacillus anthracis* is the anthrax pathogen; *Rattus norvegicus* is the Norwegian brown rat, and *Homo sapiens* is the human. The yeast, fruit fly, *C. elegans*, arabidopsis, and the rat are often used as model organisms.

system is not yet able to produce enough people with the truly multidisciplinary background and knowledge requested in areas such as bioinformatics. Hence, there is a large gap between the need to effectively analyze the mountains of data being generated continuously and the number of people able to perform such analyses in the best possible ways. Bioinformatics experts able to analyze data are and will continue to be very valuable to many employers. It is hope that this book will help you get closer to becoming such an expert.

The avalanche of data resulting from the progress in the field of molecular biology was started by understanding the nature of molecular information in living organisms and the development of new and precise high throughput screening methods. Molecular biology deals primarily with the information that macromolecules, such as the **deoxyribonucleic acid** (**DNA**) and the **ribonucleic acid** (**RNA**) carry, their interrelationship, and role in cells. Therefore, a brief description of the cell and its very basic mechanisms follows in the next chapter. Although these concepts are not absolutely necessary in order to understand the data analysis methods and techniques presented in the rest of the book, it is always better if one has a basic understanding of where the numbers analyzed come from and what they actually mean.

Chapter 2

The cell and its basic mechanisms

. . . I could exceedingly plainly perceive it to be all perforated and porous, much like a Honey-comb, but that the pores of it were not regular. . . . these pores, or ***cells****, . . . were indeed the first microscopical pores I ever saw, and perhaps, that were ever seen, for I had not met with any Writer or Person, that had made any mention of them before this. . .*

—Robert Hooke, Micrographia, 1665

2.1 The cell

The cell is the building block of all organisms. Fig. 2.1 shows a typical eukaryotic cell as well as a typical prokaryotic cell. A **eukaryotic** cell (top panel in Fig. 2.1) has a nucleus and is found in more evolved organisms. A **prokaryotic** cell (bottom panel in Fig. 2.1) does not have a nucleus and is always single-cellular, e.g., bacteria. As shown in these figures, each cell is a very complex system that includes a number of parts and structures.

The main parts of a eukaryotic cell include the **membrane**, the **cytoplasm**, the **mitochondria**, the **microtubules**, the **lysosomes**, the **ribo-**

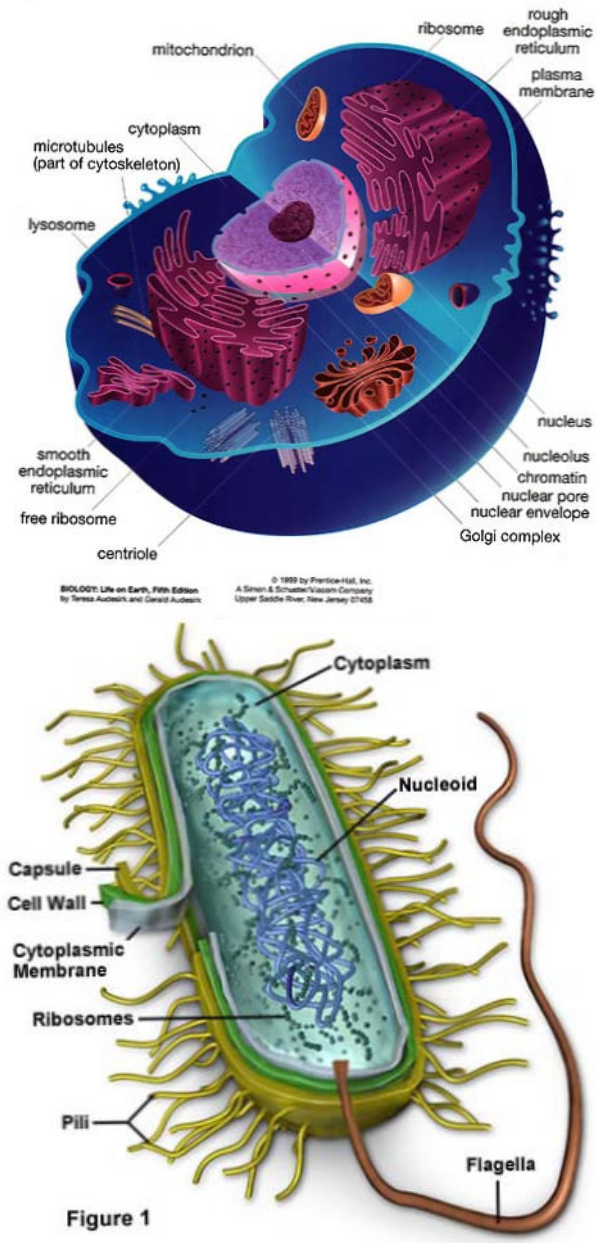


FIGURE 2.1: Two cells. A eukaryotic cell (top panel) has a nucleus and is found in more evolved organisms. A prokaryotic cell (bottom panel) does not have a nucleus and is mostly found in bacteria. The eukaryotic cell figure is from *Life on Earth*, Audesirk et al., Prentice Hall. Printed with permission. The prokaryotic cell is copyrighted Michael W. Davidson. Printed with permission.

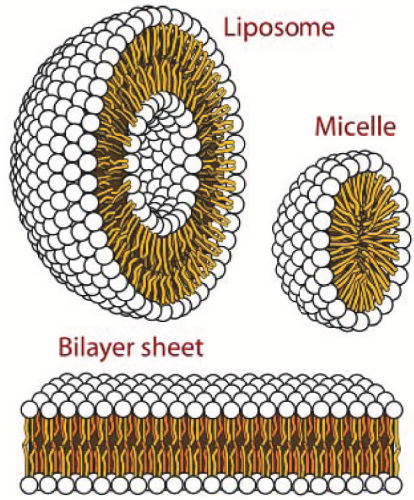


FIGURE 2.2: Cross section of the different structures that phospholipids can take in an aqueous solution. The circles are the hydrophilic heads and the wavy lines are the fatty acid side chains. The bilayer appears in the cell membrane, nuclear membrane, vesicles, etc. Author: Mariana Ruitz, released in the public domain.

somes, the **smooth** and **rough endoplasmic reticula**, etc. In a eukaryotic cell, there is also a **nucleus** that hosts a **nucleolus** and the **chromatin** within a **nuclear envelope** that features some **pores**. In the following, we will briefly discuss these.

The **cellular membrane** generally consists of two layers of phospholipid molecules. Each such molecule has a polar hydrophilic head and two non-polar (hydrophobic) tails. Since both the cytoplasm inside the cell as well as the extracellular environment contain a lot of water, the membrane molecules are aligned in a double layer, each layer presenting the hydrophilic head on the surface of the membrane (both inside and outside the cell), while all the hydrophobic tails point towards each other within the membrane. Fig. 2.2 shows the basic structure of a membrane, as well as those of two other structures that can be formed by the phospholipid bilayer. However, the cellular membrane has a structure that is far more complex than a simple phospholipid bilayer. Among other molecules, it also includes some proteins called integral membrane proteins. These can appear on the inside surface of the membrane,

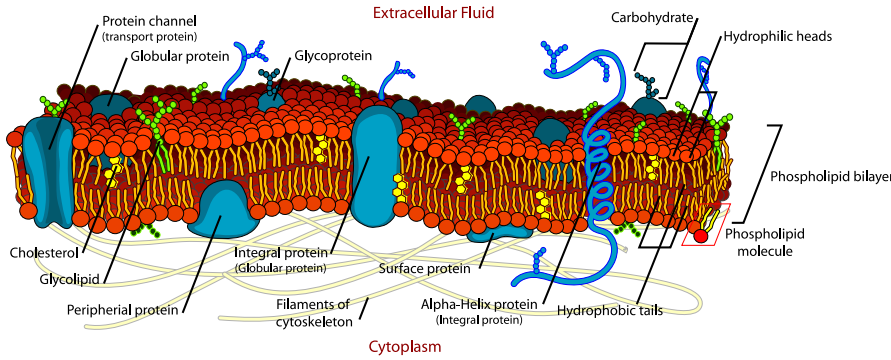


FIGURE 2.3: The cell membrane generally consists of two layers of phospholipid molecules (the basic bilayer shown in Fig. 2.2) but also has a number of other features, including surface proteins present either on the inside or on the outside surface of the membrane, integral proteins of various types crossing both layers of the membrane, channel-forming proteins acting like gateways for certain molecules, transmembrane proteins that have receptors on the outside surface and are able to trigger specific intracellular responses when their target ligand is present in the extracellular space, etc. Author: Mariana Ruitz, released in the public domain.

on its outside surface or crossing the membrane (in which case they are called transmembrane proteins). The outside part of such a protein is called a receptor. Its role is to bind to a given molecule, called ligand, when this molecule is present outside the cell. Generally, when this happens, the transmembrane protein will initiate an intracellular response. Other transmembrane proteins act as gateways, allowing certain molecules from outside the cell to enter the cell through a channel formed by the protein. The Fig. 2.3 shows a crosscut through the cellular membrane illustrating some of these additional features of the cellular membrane. The membrane is involved in several very important processes such as cell adhesion, cell signaling, and ion channel conductance.

The **cytoplasm** includes everything that is in the cell (organelles, water, other chemical molecules, etc.), except the nucleus. The **nucleoplasm** includes everything that is in the nucleus. Together, the cytoplasm and the nucleoplasm form the **protoplasm**.

Organelles are specialized sub-cellular structures of the cytoplasm. In some sense, the organelles do for the cell what the organs do for complex organisms: each organelle has a very specific function in the complex mechanisms that keep the cell alive. Some authors define the organelles as being membrane-bound structures that have some specific function in the cell. Other authors use a less restrictive definition considering that any structure that carries out a particular and specialized function is an organelle, whether it is

membrane-bound or not. For instance, the ribosome (see details below) is an organelle according to the latter definition, but not according to the former.

The **chromatin** is a structure made of proteins and highly packed DNA. The DNA contains the **genes** that code for all proteins, as well as other functional and control elements. A prokaryotic cell does not have a nucleus, and the DNA material is found directly in the cytoplasm.

The **mitochondria** (singular **mitocondrion**) are the power plants of the cell (see the eukaryotic cell in Fig. 2.1). Their main role is to produce energy for the cell. There are several fascinating facts about mitochondria. First, the mitochondria have their own DNA (called **mitochondrial DNA** or **mtDNA**), which is circular, very much like the DNA of a bacteria. Based on this as well as other data, it has been suggested that the mitochondria are in fact what is left from a small prokaryote cell that was swallowed millions of years ago by an eukaryote, or perhaps a larger prokaryote [299]. Rather than digesting the smaller prokaryote, the bigger cell found out that a symbiotic partnership would be much better for both of them. The smaller cell benefits from the free basic fuel, as well as the safe environment provided by the larger cell. In turn, the larger cell benefits from the energy produced by the smaller cell. This is an example of mutualism, a type of symbiosis in which both organisms benefit and neither is harmed. This partnership is so strong now that eukaryotic cells cannot survive without mitochondria, and the **endosymbionts** (the smaller cells which were incorporated in the eukaryotic cell) also cannot survive on their own. The same mutualism is found in plants and algae whose cells contain chloroplasts, organelles able to transform sunlight into energy during the process of photosynthesis.

The second fascinating fact about mitochondria is that in most multicellular organisms (including human), the mitochondrial DNA is inherited from mother to child. This is unlike the DNA in the nucleus, which is formed in the offspring by combining the nucleic DNA from both mother and father. In fact, this very unusual property of the mitochondrial DNA is at the center of a book by Bryan Sykes, titled *The Seven Daughters of Eve.* In this book, the author describes how the entire population of Europe can be traced back to only seven women (hence the title) using this property of the mtDNA. In fact, in the same book, Bryan Sykes uses the same argument to refute Thor Heyerdahl's hypothesis that the population of Polynesia originated in South America and reached the Polynesian islands by crossing the Pacific on primitive bamboo rafts.[1]

[1]This hypothesis may seem rather far-fetched, but actually in 1947, Thor Heyerdahl built such a bamboo raft and sailed it more than 4,300 miles (8,000 Kms) from South America to the Tuamotu Islands. The raft, Kon-Tiki, was built using exclusively materials and technologies available at a time to those populations. For instance, since those materials and technologies did not include either iron or nails, the logs of the raft were held together with manually weaved hemp ropes [207]. Even though his theory regarding the origins of the Polynesian population was ultimately proved incorrect, Thor Heyerdahl will remain in history for the courage and determination to put his life on the line in order to prove the feasibility of his scientific hypothesis.

Since the mitochondria have their own DNA, it follows that they can reproduce independently of the host cell. Thus, the number of mitochondria can vary in time within a given cell or cell type. Various treatments such as some antifungal treatments can actually kill many mitochondria. After the treatment is stopped, the mitochondria will start multiplying again and their population will eventually recover to normal levels. Recently, it has been discovered that mitochondria also play an essential role in mechanisms other than energy generation. For instance, mitochondria can be the target of the immune system response [301]. Thus, it has recently been shown that the killer T cells of the immune system can trigger the programmed cell death (or apoptosis) of virus-infected or cancer cells by releasing two serine proteases called granzymes in the target cells. One of these, Granzyme A, targets a certain protein NDUFS3 which is degraded. In turn, this causes mitochondria to produce damaging reactive oxygen which eventually causes cell death. The other granzyme, Granzyme B, causes the breakdown of the outer mitochondrial membrane which releases a number of death-promoting proteins activating a chain-reaction known as the caspase protease cascade resulting in massive DNA damage and cell death.

The **microtubules** are cylindrical hollow structures with a diameter of approximately 25 nm and length varying from 200 nanometers to 25 micrometers, which are found in the cytoplasm of eukaryotic cells (see Fig. 2.4). They are part of the cytoskeleton of the cell, providing structural support (e.g., giving the cell its shape), and assisting in cellular locomotion, transport, and cell division.

The **lysosomes** are vesicles that contain digestive enzymes called acid hydrolases. The term lysosome comes from *lysis* (dissolution or destruction in Greek) and *soma* (body in Greek). The lysosome is involved in the breakdown of various materials such as food particles, viruses, or bacteria that managed to penetrate into the cell. Because the lysosome membrane isolates the contents of the lysosome from the rest of the cell, the inside of the lysosome can be maintained at a pH that is much more acidic (around 4.5–4.8) than the neutral pH (7.0) of the intracellular fluid, or **cytosol**. This acidic pH is necessary for the functioning of the hydrolases. This plays the role of a safety mechanism since if the digestive enzymes are released in the cell by accident they will not harm it, as long as the pH in the cytosol is normal. However, if the cell is dead, dying, or injured, the acid hydrolases released from lysosomes can self-digest the cell in a process of **autolysis**. This is why sometimes lysosomes are also called suicide bags.

The **ribosomes** are complexes of RNA and proteins whose main role is to translate messenger RNA (mRNA) into chains of polypeptides using amino acids delivered by the transfer RNA (tRNA) molecules. The term ribosome comes from *ribo*nucleic acid (RNA) and *soma* (body in Greek). The role of ribosomes will be discussed in more detail in Section 2.3.

The **endoplasmic reticulum** (plural endoplasmic reticula) (ER) is an interconnected structure composed of **tubules**, **vesicles**, and **cisternae**. A

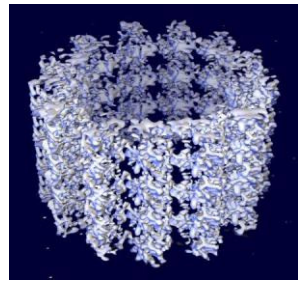


FIGURE 2.4: A microtubule is a hollow cylindrical structure made out of the α and β tubulin proteins. The microtubules have many roles, including providing structural support, assisting in cellular locomotion, and cell division. Image in the public domain.

tubule is a small tube-like structure. A vesicle is a small sac surrounded by a membrane similar to the cellular membrane. Vesicles store, transport, or digest various cellular products or waste. A cisterna (plural cisternae) is a flattened disc surrounded by a membrane. They also carry proteins. The ERs are involved in the translation of certain specialized proteins, the transport of proteins to be used in the cell membrane or to be secreted from the cell, sequestration of calcium, and production and storage of certain macromolecules. The ERs can be subdivided into **rough endoplasmic reticula**, **smooth endoplasmic reticula**, and **sarcoplasmic reticula**, each having some slightly different characteristics and roles.

Fig. 2.5 shows the endomembrane system (from *endo* meaning internal in Greek, and membrane) in a eukaryotic cell, including the rough and smooth endoplasmic reticula, secretory vesicles, lysosomes, the Golgi apparatus, etc.

The **nucleus** (plural **nuclei**) is the largest organelle in the cell. In mammalian cells, the nucleus measures about 11–22 micrometers in diameter and occupies about 10% of its volume. The nucleus contains most of the cell's nuclear genetic material. This is in the form of very long linear DNA molecules organized most of the time into a DNA-protein complex structure called chromatin, which is essentially very tightly packed double-stranded DNA. During cell division, the chromatin forms well-defined structures called chromosomes. Each chromosome contains many genes as well as long sequences of intergenic DNA. The main roles of the nucleus are to protect the nucleic DNA, to control

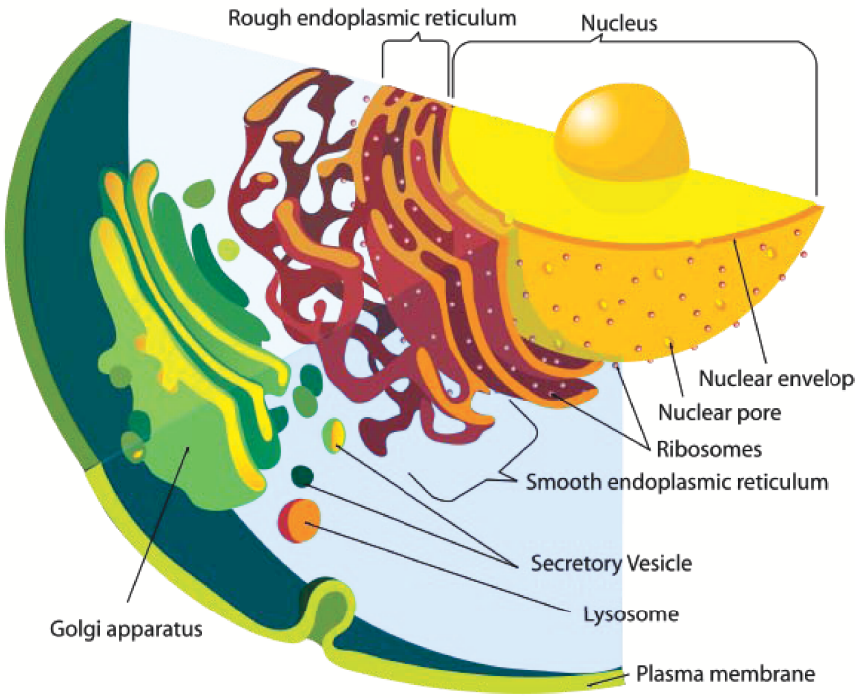


FIGURE 2.5: The endomembrane system is a system of intracellular membranes that divide the eukaryotic cells into various organelles. The endomembrane system includes the cell membrane itself, the nuclear envelope that separates the nucleus from the cytoplasm, the smooth and rough endoplasmic reticula, the Golgi apparatus, the lysosomes, vesicles, etc.

the gene expression process, and to mediate the replication of DNA during the cell cycle.

The nucleus is surrounded by a nuclear membrane that separates it from the cytoplasm. Since this membrane is impenetrable to most molecules, it has a number of small orifices, or **pores**, which allow certain small, water-soluble molecules to penetrate the nuclear membrane in very specific conditions. Larger molecules such as proteins must be transported in a very carefully controlled way by specialized transporter proteins. The surface of the nucleus is also studded with ribosomes much like the surface of the rough endoplasmic reticulum which continues it. Although the interior of the nucleus is not separated by other membranes, its content is not uniform. Fig. 2.6 shows that the nucleus has a central part called nucleolus, which is mainly involved in the assembly of ribosomes, and two types of chromatin: **heterochromatin** and **euchromatin**. The euchromatin is the less dense of the two and contains those genes that are expressed often by the cell. The structure of the euchromatic resembles that of a set of beads on a string (see Fig. 2.7 and Fig. 2.8). The heterochromatin is the more compact form and contains genes that are transcribed only infrequently, as well as chromosome constitutive elements such as **telomeres** (repetitive DNA that appears at the end of the chromosomes protecting them from destruction) and **centromeres** (the central region of a chromosome, where the arms of the chromosome are joined together).

2.2 The building blocks of genomic information

2.2.1 The deoxyribonucleic acid (DNA)

DNA is most commonly recognized as two paired chains of chemical bases, spiraled into what is commonly known as the double helix. DNA is a large polymer with a linear backbone of alternating sugar and phosphate residues. The sugar in DNA molecules is a 5 carbon sugar (deoxyribose); successive sugar residues are linked by strong (covalent) phosphodiester bonds. A nitrogenous base is covalently attached to carbon atom number $1'$ (one prime) of each sugar residue. There are four different kinds of bases in DNA, and this why it simple to understand its basic function and structure. The order in which the bases occur determines the information stored in the region of DNA being looked at.

The four types of bases in DNA are adenine (A), cytosine (C), guanine (G), and thymine (T) each consisting of heterocyclic rings of carbon and nitrogenous atoms. The bases are divided into two classes: purines (A and G) and pyrimidines (C and T). When a base is attached to a sugar, we speak of a nucleoside. If a phosphate group is attached to this nucleoside, then it becomes a nucleotide. The nucleotide is the basic repeat unit of a DNA strand.

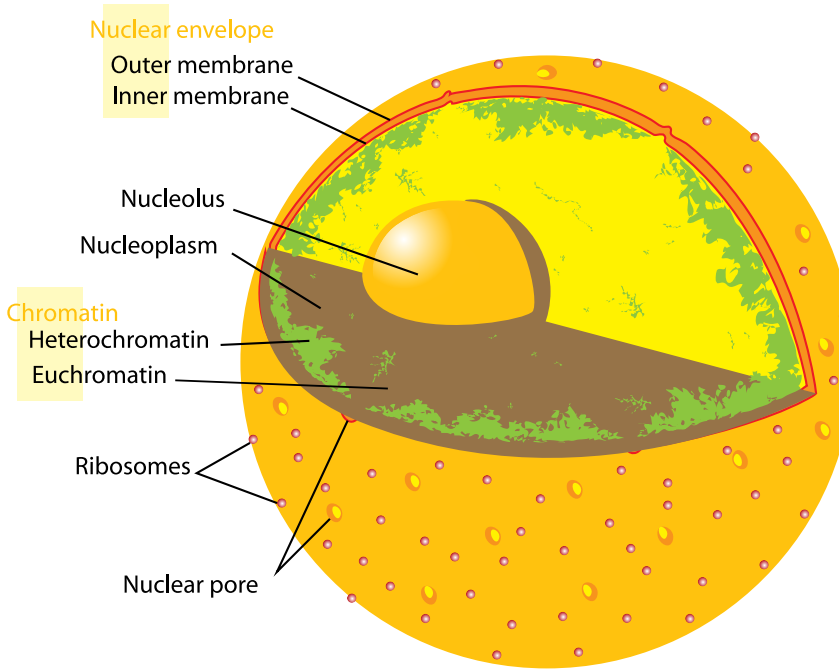


FIGURE 2.6: The nucleus is the largest organelle in a cell. The nucleus has a central part called nucleolus, which is mainly involved in the assembly or ribosomes, and two types of chromatin: heterochromatin and euchromatin. The nucleus is surrounded by a nuclear membrane that separates it from the rest of the cytoplasm. This membrane is studded with pores and ribosomes. The pores allow certain small, water-soluble molecules to penetrate the nuclear membrane in very specific conditions. Author: Mariana Ruiz, released in the public domain.

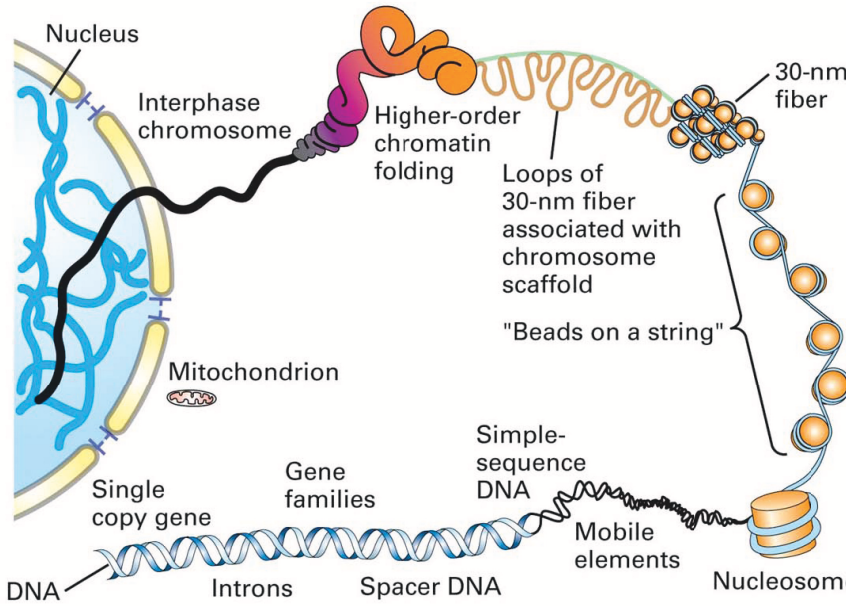


FIGURE 2.7: The DNA material in the nucleus is tightly packed in a complex way. The very long double-stranded DNA that contains the genes is sometimes compared with a string. From place to place along this string, there are cylindrical structures called **histones**. The scale is such that if the DNA is compared with a string, the histones could be compared with some beads on this string, only that instead of the string going through each bead, the string is wrapped around each bead. The double-stranded DNA wrapped around a histone forms a **nucleosome**. The "beads on a string" structure can be further folded and packed even tighter in loops of DNA fiber that are further folded and compacted to form the chromatin. In order to be transcribed, the DNA encoded for a gene needs to be accessible so the location of a gene in relationship with the histones and other structures may be important for the gene expression process. Figure from *Molecular Cell Biology*, 5th Edition (2004), Lodish, H., et al., printed with permission.

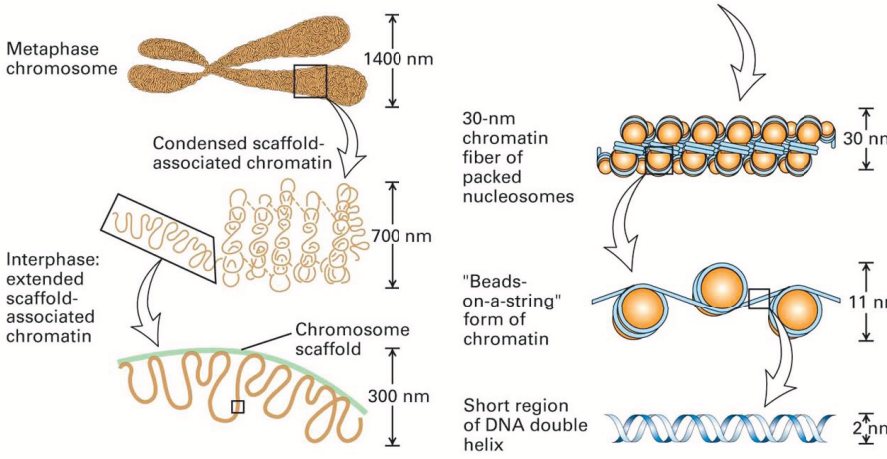


FIGURE 2.8: The scales of various structures used in chromatin packing. Figure from *Molecular Cell Biology*, 5th Edition (2004), Lodish, H., et al., printed with permission.

The formation of the double helix is due to the hydrogen bonding that occurs between laterally opposed bases. Two bases form a **base pair** (bp). The chemical structure of the bases is such that adenine (A) specifically binds to thymine (T) and cytosine (C) specifically binds to guanine (G). These are the so called Watson-Crick rules. Since no other interactions are possible between any other combination of base pairs, it is said that A is complementary to T and C is complementary to G.[2] Two strands are called complementary if, for any base on one strand, the other strand contains this base's complement. Two complementary single-stranded DNA chains that come into close proximity react to form a stable double helix (see Fig. 2.9) in a process known as **hybridization** or **annealing**. Conversely, a double-stranded DNA can be split into two complementary, single-stranded chains in a process called **denaturation** or **melting**. Hybridization and denaturation play an extremely important role both in the natural processes that happen in the living cells and in the laboratory techniques used in genomics. Because of the base complementarity, the base composition of a double-stranded DNA is not random. The amount of A equals the amount of T, and the amount of C is the same as the amount of G.

Let us look at the backbone of the DNA strand again. Phosphodiester

[2]In fact, other interactions are possible but the A-T and C-G are the ones that occur normally in the hybridization of two strands of DNA. This is because the A-T and C-G pairings are the ones that introduce a minimal distortion to the geometrical orientation of the backbones.

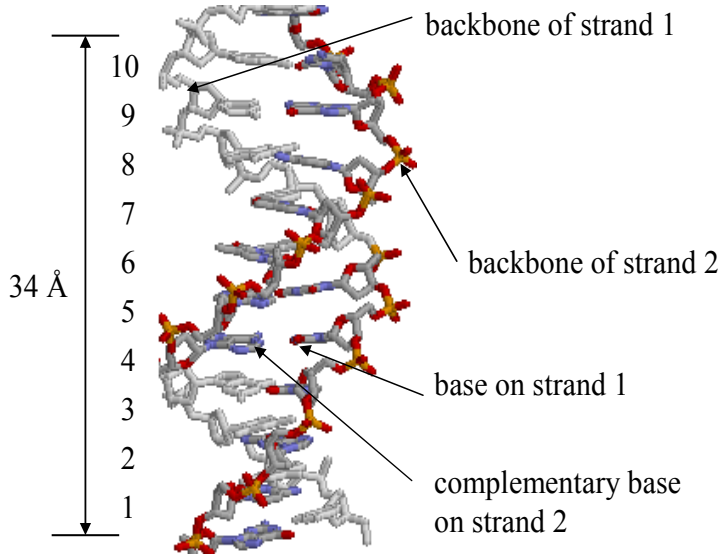


FIGURE 2.9: A short fragment (10 base pairs) of double-stranded DNA. Image obtained with Protein Explorer.

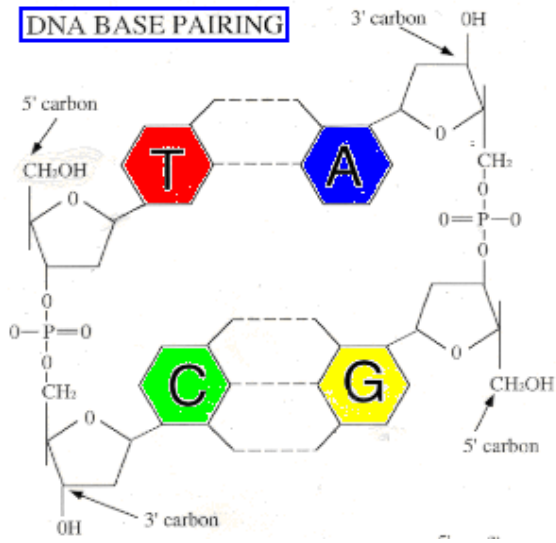


FIGURE 2.10: Each end of a single strand of DNA is identified by the carbon atom that terminates the strand (5′ or 3′). Two strands of DNA always associate or anneal in such a way that the 5′ → 3′ direction of one DNA strand is the opposite to that of its partner.

bonds link carbon atoms number 3′ and 5′ of successive sugar residues. This means that in the terminal sugar the 5′ is not linked to a neighboring sugar residue. The other end is termed 3′ end, and it is characterized by the lack of a phosphodiester bond on that particular carbon atom. This gives a unique direction to any DNA strand. By convention, the DNA strand is said to run from the 5′ end to the 3′ end. The two strands of a DNA duplex are considered to be antiparallel. They always associate or anneal in such a way that the 5′ → 3′ direction of one DNA strand is the opposite to that of its partner (see Fig. 2.10).

The two uprights of the DNA ladder are a structural backbone, supporting the rungs of the ladder. These are also the information-carrying parts of the DNA molecule. Each rung of the ladder is made up of two bases that are paired together. This is what makes the steps of the spiral staircase. The two-paired bases are called a base pair as described earlier. The length of any DNA fragment is measured in base pairs (bp), similarly to how we measure length in inches. However, since the DNA is formed with base pairs, the length of a DNA fragment can only be a discrete number of such pairs, unlike a length which can include fractions of an inch.

Each nucleotide has a discrete identity. The sequence of the nucleotides in a DNA can be read by the "machinery" inside the cell. Genes, which represent

large sequences of DNA, can be looked at as instructions telling the cell how much protein to make, when it should be made, and the sequence that can be used to make it. The information in the DNA is like a library. In the library, you will find books, and they can be read and reread many times, but they are never used up or given away. They are retained for further use. Similarly, the information in each gene is read (see below), perhaps millions of times in the life of an organism, but the DNA itself is never used up.

2.2.2 The DNA as a language

Each base can be thought of as a specific letter of a 4-letter alphabet, combining to form words and sentences. In essence, a gene is a recipe for making a protein. Let us consider for example the following very simple recipe for making an omelette:

Take three eggs; scramble; add oil in a pan; heat it up; add the eggs; get the omelette.

From a syntactic perspective, this is a simple string of characters from a set that includes the 52 alphabetic characters in the English language (lower and upper case) plus some special characters (space, semicolon, period, etc.). Usual grammatical conventions tells us to start a phrase using a capital letter and end it using a period. If we were to be minimalistic and restrict ourselves to the 26 lower-case characters, we can give up the spaces and be explicit about the initial capital letter and the final period. In this case, we could write the recipe as:

capitaltakethreeeggsscraambleaddoilinapanheatitupaddtheeggsgettheomelettepe
riod

In the coding above, the "capital" and "period" markers are there just to indicate to us when a recipe starts and ends, in case we want to put together a collection including many such recipes. Similarly, since one chromosome contains many genes in a unique, very long DNA sequence, the beginning and end of a gene are indicated by special markers called **start** and **stop codons**.

Furthermore, in a digital computer system, the recipe above would be stored in a binary format, usually using 8 bits (or one byte) for every alphanumeric and special character. The recipe above would now look something like this:

01100011	01100001	01110000	01101001	01110100	01100001	...
c	a	p	i	t	a	...

From this, it follows that any time such a recipe is accessed in the memory of a computer, a translation has to take place from the binary alphabet {0,1} used by the computer to the English alphabet used by humans. Similarly, since a protein is a sequence of amino acids, and its recipe is stored as a sequence of

DNA bases, every time a protein is produced a translation has to take place that would map the recipe written in the 4-letter alphabet of the DNA, to the necessary protein sequence that uses amino acids from a 20 letter alphabet.

The DNA bases are grouped in triplets, or **codons**, for the same reason bits are grouped in octets, or bytes: if each symbol is limited to only two values (in the binary alphabet), or four values (in the DNA alphabet), groups of several symbols are needed in order to represent symbols from larger alphabets such as the set of 20 amino acids (for proteins) or the set of 26 letters of the English alphabet (for text). In fact, soon after the discovery of the DNA, the existence of a three-letter code to map from the DNA alphabet to the amino acid alphabet was postulated by George Gamov based on the fact that n=3 is the smallest value of n that satisfies $4^n > 20$. In other words, n=3 is the smallest size of a tuple for which there are more tuples than the 20 amino acids that needed to be coded for.

Each triplet of DNA nucleotides, or each codon, corresponds to a certain amino acid. Fig. 2.11 shows the correspondence between all possible codons and their respective amino acids, as well as some structural information, chemical properties and post-translational modifications of the various amino acids. The mapping from codons to amino acids is known as the genetic code. There is a start codon that indicates where the translation should start and several end codons that indicate the end of a coding sequence. Note that since there are $4^3 = 64$ different codons and only 20 different amino acids, it follows that either several different codons have to code for the same amino acid or many codons have to code for no amino acids at all. In fact, the former is true, with most amino acids being coded for by more than one codon. This may be more easily visible in Fig. 2.12. For instance, if a codon has the first two nucleotides C and T[3], the codon will be translated into leucine independently of the third nucleotide. Similarly, the CC* codon will be translated into proline, etc.

It turns out that there is another level of complexity about the genetic code. At the same time with the protein coding information, the genome has to also carry other types of signals such as regulatory signals telling the cell when to start and stop protein production for each protein, signals for splicing, etc. It has been shown recently that when these additional requirements are taken into consideration, the universal genetic code that we know is nearly optimal with respect to all other possible codes [15]. The optimality here was defined as the ability to minimize the effects of the most disastrous type of errors, the frame-shifts, as well as the property that close codons (codons that differ by only one letter) are mapped to either the same amino-acid or to chemically related ones. This means that, if a translation process misreads a single letter, the error introduced will have no or little consequences. In the same context, it has also been noted that amino acids with a simple chemical structure tend to have more codons assigned to them [177].

[3]For reasons that will be explained soon, the CT tuple appears as CU in Fig. 2.11.

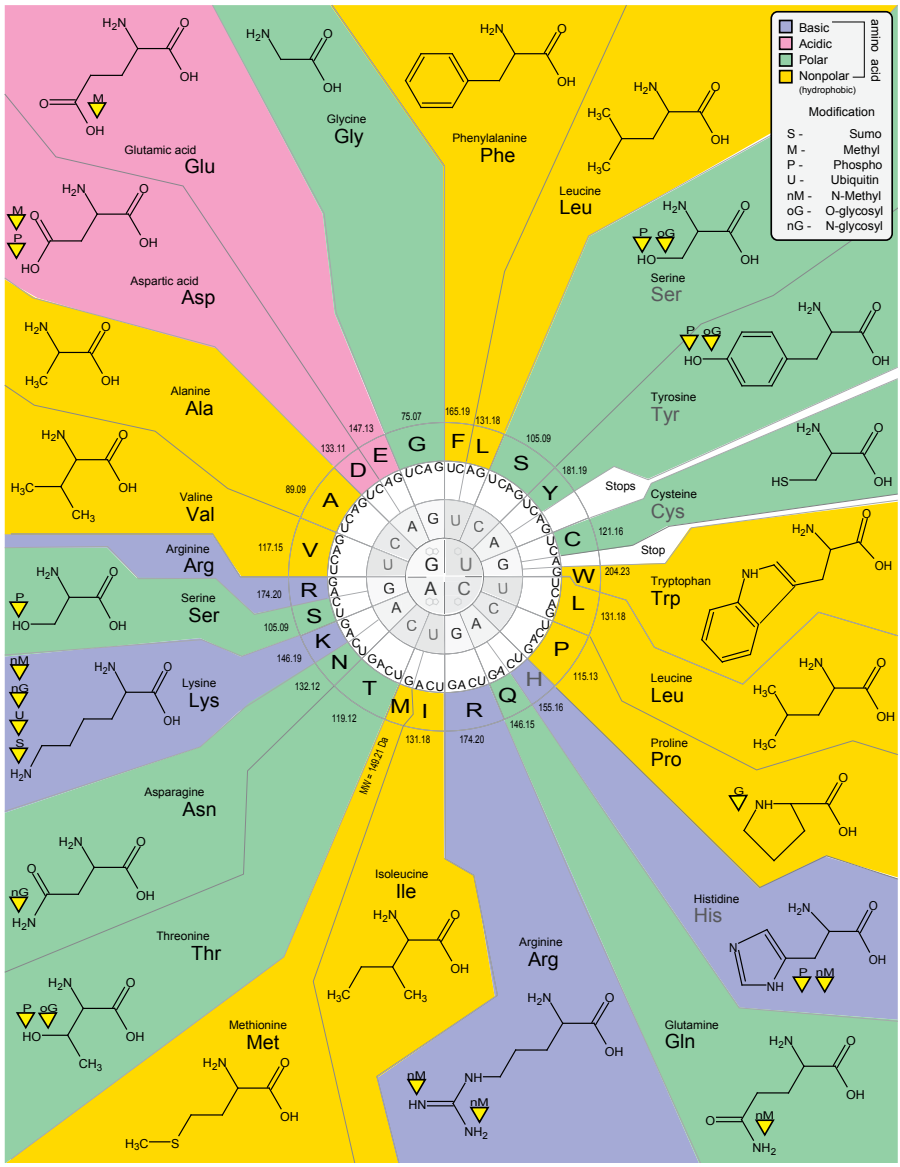


FIGURE 2.11: The genetic code is the mapping between the 4 letter-alphabet of the DNA/RNA nucleotides that appear in the genes, and the 20 letter-alphabet of the amino acids that form the proteins coded for by the genes. In this image, one starts in the middle with the first nucleotide of a codon and goes outwards following the remaining nucleotides. The figure also shows some structural information, chemical properties, as well as various possible post-translational modifications. Original image in public domain by Kosi Gramatikoff courtesy of Abgent; modified by Seth Miller and the author.

First Position	Second Position: T	Second Position: C	Second Position: A	Second Position: G	Third Position
T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T
	TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C
	TTA Leu [L]	TCA Ser [S]	TAA *Ter* [end]	TGA *Ter* [end]	A
	TTG Leu [L]	TCG Ser [S]	TAG *Ter* [end]	TGG Trp [W]	G
C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T
	CTC Leu [L]	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C
	CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A
	CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	G
A	ATT Ile [I]	ACT Thr [T]	AAT Asn [M]	AGT Ser [S]	T
	ATC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C
	ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A
	ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G
G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	T
	GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C
	GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A
	GTG Val [V]	GCG Alo [A]	GAG Glu [E]	GGG Gly [G]	G

FIGURE 2.12: Another view of the genetic code. In this figure, it may be easier to see the redundancy intrinsic to the code. For instance, if a codon has the first two nucleotides C and T, the codon will be translated into leucine independently of the third nucleotide. Similarly, the CC* codon will be translated into proline, etc. There are 3 stop codons that mark the end of a coding sequence ("*end*" above) and one start codon marking the begining of a coding sequence ("*M*" above).

2.2.3 Errors in the DNA language

The DNA sequence of a chromosome can also contain errors. There are several types of errors as follows:

1. **Mutations**. A mutation is an situation in which one nucleotide is substituted by another one. For instance, in our original recipe:

   ...takeeggsscrambleaddoil...

   a mutation could substitute the second "g" with an "o" to yield:

   ...takeegosscrambleaddoil...

   leading to a rather different result:

   ...take egos scramble add oil...

   Mutations as above are also known as single point mutations or single nucleotide mutations. Since the genetic code is such that different codons can correspond to the same amino acid, a single point mutation in the DNA sequence of a gene may not change at all the structure of the protein coded for by the given protein. However, it is also possible that a single point mutation is extremely disruptive or even lethal for a given organism.

2. **Deletions**. A deletion is a situation in which one of the nucleotides is missing. For instance, in our original recipe:

   ...takeeggsscrambleaddoilinapanheatitupaddegsandgetomellette ...

   a deletion could remove the highlighted "h" to yield:

   ...takeeggsscrambleaddoilinapaneatitupaddegsandgetomellette...

   leading again to a departure from the intended outcome, in this case, an early consumption of the rather raw dish:

   ...take eggs, scramble, add oil in a pan, eat it up...

3. **Insertions**. An insertion is a situation in which an additional nucleotide is inserted in the middle of an existing sequence. For instance, in our original recipe:

   ...takeeggsscrambleaddoil...

   an insertion could add the highlighted "s" to yield:

   ...takeeggsscrambleaddsoil...

   leading to the consumption of something rather different from the intended omelette:

   ...take eggs scramble add soil...

4. **Frame shifts**. A frame shift is a situation in which the DNA sequence is shifted. Such a shift can be caused for instance by an insertion or

deletion. Because of this shift, the identity of all subsequent codons will be changed. For instance, in our original binary-coded recipe above:

01100011	01100001	01110000	01101001	01110100	01100001	...
c	a	p	i	t	a	...

a frame shift can cause the reading to start at the second digit instead of the first one, which will mean that the same bits will be grouped, and hence interpreted, very differently:

...0	11000110	11000010	11100000	11010010	11101000	11000010
	ä	-	Ó	Ê	ч	–

In this example, a frame shift mutation that moved the start point by just one position to the right, caused changes so substantial in the subsequent bytes that all the characters became special characters and the text fragment itself became completely unintelligible. Similarly, most frame shift mutations in a DNA sequence will cause very substantial changes in the corresponding amino acid sequence. In fact, this is how the organization of the protein-coding DNA in triplets was originally proven. In a 1961 experiment, Crick, Brenner et. al. [105] showed that either inserting or deleting only one or two nucleotides prevents the production of a functional protein while inserting or deleting 3 nucleotides at a time still allows the functional protein to be produced.

Together, DNA works with protein and RNA (ribonucleic acid) in a manner that is similar to the way a computer works with programs and data. DNA has a code that is continuously active, having instructions and commands such as "if-then," "go to," and "stop" statements. It is involved with regulating protein levels, maintaining quality control, and providing a database for making proteins. The best analogy to use when describing DNA and its function in cells is to look at a cell as a little factory. In a cell, much like in a factory, specific items are produced in specific places. There is a certain flow of materials through the cell, and there are various communication and feedback mechanisms regulating the speed with which various processes happen in various parts of the cell. Such communication and feedback mechanisms also allow the cell to adapt the speed of its internal processes to the demands of the environment much like a factory can adjust its production in response to the changing needs of the market.

2.2.4 Other useful concepts

A **gene** is a segment or region of DNA that encodes specific instructions, that allow a cell to produce a specific product. This product is typically a protein,

such as an enzyme. There are many different types of proteins. Proteins are used to support the cell structure, to break down chemicals, to build new chemicals, to transport items, and to regulate production. Every human being has about 22,000 putative genes[4] that produce proteins. Many of these genes are totally identical from person to person, but others show variation in different people. The genes determine hair color, eye color, sex, personality, and many other traits that in combination make everyone a unique entity. Some genes are involved in our growth and development. These genes act as tiny switches that direct the specific sequence of events that are necessary to create a human being. They affect every part of our physical and biochemical systems, acting in a cascade of events, turning on and off the expression, or production, of key proteins that are involved in the different steps of development.

The key term in growth and development is **differentiation**. Differentiation involves the act of a cell changing from one type of cell, when dividing through mitosis, into two different types of cells. The most common cells used to study differentiation are the stem cells. These are considered to be the "mother cells," and are thought to be capable of differentiating into any type of cell. It is important to understand the value of differentiation when learning about genetics and organism development. Everyone starts life as one single cell, which divides several times before any differentiation takes place. Then, in a specific stage, differentiation begins to take place, and internal organ cells, skin cells, muscle cells, blood cells, etc. are created.

Another very important process is the DNA **replication**. The DNA replication is the process of copying a double-stranded DNA molecule to create two identical double-stranded molecules. The original DNA strand is called the template DNA. Each strand of the double-stranded template DNA becomes half of a new DNA double helix. Because of this, the DNA replication is said to be **semi-conservative**. This process, illustrated in Fig. 2.13, must take place before a cell division can occur.

The replication process starts at some fixed locations on the DNA strand called **replication origins**. These replication origins have certain features such as repeats of a short sequence that create a pattern that is recognized by an **initiation protein**. Once this protein binds to the double strand, the template DNA is unwound and separated into the two component single strands (see the topoisomerase and the helicase in Fig. 2.13). This region is called the **replication fork** and travels along the DNA sequence as the replication takes place. The single-stranded regions are trapped by **single-stranded binding proteins** making them accessible to the DNA polymerase. The DNA polymerase is an enzyme that can add single nucleotides to the $3'$ end of an existing single strand by matching the nucleotides in the template. However, the DNA polymerase needs an existing sequence with a free $3'$ end before it can do its job. The short nucleotide sequence that provides the starting point for the

[4]Initially, the number of human genes was estimated at 100,000 to 140,000. Subsequently, the estimate was revised down to about 20,000 to 30,000. At the moment of this writing, there still exists a controversy over this number [340].

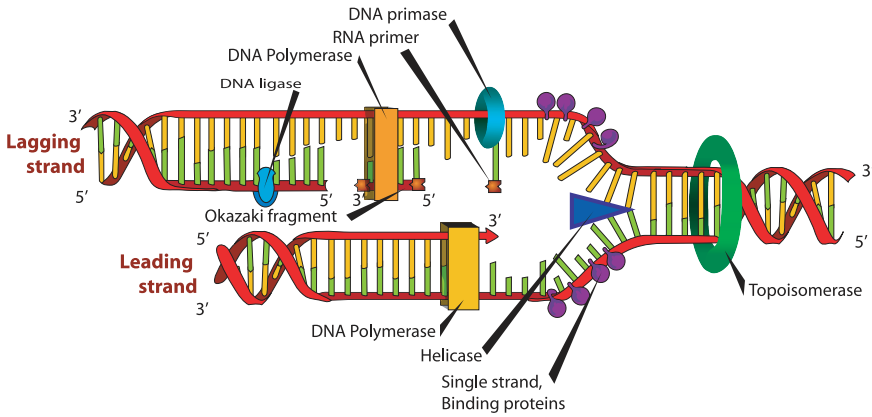


FIGURE 2.13: The main processes involved in the DNA replication. The topoisomerase is an enzyme that facilitates the unwinding of the double-stranded DNA, allowing the helicase enzyme to separate the two DNA strands. These single strands are kept apart by certain single-strand binding proteins allowing the DNA polymerase and the other enzymes to have access to the single strands. The construction of the new strands starts from primers built by a primase enzyme. On the lower strand, called the **leading strand**, the newly constructed strand is continuously elongated by the DNA polymerase that adds new nucleotides to the free 3′ end of the new strand (left to right in this figure), replicating the template. On the upper strand, called the **lagging strand**, the new strand is elongated intermittently by adding new Okazaki fragments which are continuously constructed by The gaps are repaired by a DNA ligase enzyme. Image in public domain by Mariana Ruiz, modified by the author.

DNA polymerase is called a **primer**. During the cell replication, the primer is constructed *ex novo* by an enzyme called **primase**. The primase actually synthesizes an RNA primer that will be later removed by an enzyme called **RNase H**, allowing the polymerase to add the correct DNA nucleotides instead.

As stated above, the DNA polymerase can only add nucleotides to the 3′ end of a strand. Hence, it follows that a DNA strand can only be constructed in one direction, from the 3′ to the 5′ end. However, an examination of Fig. 2.13 shows that apparently both new strands grow at the same time. The new strand being constructed in the lower part of the figure grows from the 5′ end to the 3′ end with new nucleotides being added, as expected to the free 3′ end of the newly formed strand. However, the strand being constructed in the upper part of the figure appears to grow from the 3′ end by extending its 5′ end which we know the polymerase cannot do. In fact, this strand does grow by extending its 5′ end, but the job is not done by the polymerase. The polymerase only constructs short fragments of DNA, still going right to left, by extending the free 3′ end. These short fragments are called **Okazaki fragments** after the Japanese scientists who discovered them in 1968. These short fragments attach themselves to the template strand as dictated by the complementarity rules. Finally, the **DNA ligase**, an enzyme that can suture or ligate, gaps in double-stranded DNA comes and adds the missing link in the newly constructed strand. In essence, the lower strand in Fig. 2.13 is being elongated continuously, seamlessly by the DNA polymerase, while the upper strand is elongated by concatenating short Okazaki fragments.

Another important concept related to genes is that of a **single nucleotide polymorphism** or SNP (pronounced "snip"). A SNP is a single nucleotide difference between different individuals of the same species, or between paired chromosomes of an individual. For instance, if the DNA sequence of most individuals of a species at a certain location reads ACGTTACG while another specific individual has the sequence ACGT**A**ACG at the same location, it is said that this individual has a SNP at that particular location. In essence, a SNP is a mutation that is more widespread in the population (and perhaps not very harmful). The two alternatives associated to a SNP location are called **alleles**. Sometimes a combination of alleles at multiple loci are transmitted together from one generation to another. Such a combination of alleles is called a **haplotype**, a term obtained from contracting words "haploid genotye." The set of all alleles of an individual, in other words the entire set of differences characteristic to this individual, is called the **genotype** of that individual. The term genotype is also used to refer to a subset of genetic traits, sometimes a single such trait. In contrast, the **phenotype** is one (or more) observable characteristic(s) of an organism. In general, the phenotype is determined by the genotype as well as environmental and developmental conditions. However, there are certain phenotypes that are determined directly by the genotype, sometime by a single gene.

Organisms such as *Homo sapiens*, which are **diploid**, that is their genomes

contain two copies of each chromosome (with the exception of the sex chromosomes), can also have SNPs between the two copies of a given chromosome. These are also referred to as alleles. Most SNPs have only two alleles. The allele that is less frequent is called the **minor allele**.

If a phenotype is determined by a single gene and this gene has, let us say, two alleles, sometime the presence of one of these in any one of the two copies of the chromosome determines the phenotype independently of the allele present on the other copy of the chromosome. In this case, this allele is called the **dominant** allele. By convention, dominant alleles are written in uppercase letters, and recessive alleles in lowercase letters. For instance, A is dominant to a (and a is **recessive** to A), which means that two individuals with the AA and Aa genotypes have the same phenotype, while a third individual with the aa genotype has a different phenotype. There are also other, more complex mechanisms of inheritance.

2.3 Expression of genetic information

Genes make up only a subset of the entire amount of DNA in a cell. The human genome, for instance, contains approximatively 3.1 billion base pairs. However, less that 2% of the genome codes for proteins. Active research in the past 20 years has identified signals (specific sequences of bases) that delimit the beginning and ending of genes. These signaling areas can be thought of as regulatory elements. They are used to control the production of protein, and work as a biofeedback system. They are usually located near the beginning of a sequence that is used to code for a protein. Through protein-DNA interactions they are used to turn on and off the production of proteins. Noncoding DNA conveys no known protein-coding or regulatory information and makes up the bulk of the DNA in our cells. These intergenic regions can include highly repetitive sequences. Although this DNA is sometimes called "junk DNA," several functions have been proposed for it, including playing a role in reshaping and rearranging the genes in the genome, acting as a buffer to decrease the damage introduced by random mutations, or acting as a spacer such that genes that are transcribed often are accessible to the DNA polymerase and the other enzymes involved in transcription after all the twists, turns, and folding of the double-stranded DNA into the chromatin structure.

The flow of genetic information is from DNA to RNA to proteins. This one-way process is the expression of genetic information in all cells and has been described as the **central dogma** of molecular biology.

To make products from a gene, the information in the DNA is first copied, base for base, into a similar kind of information carrier, called a **transcript**, or **messenger RNA** (mRNA). The RNA copy of the gene sequence acts as a messenger, taking information from the nucleus (where the DNA is found in its

chromosomal form) and transporting it into the cytoplasm of the cell (where the machinery for making gene products is found). Once in the cytoplasm, the messenger RNA is translated into the product of the gene, a protein. The sequence of the protein is defined by the original sequence of the DNA bases found in the gene.

DNA is the hereditary material in all present-day cells. However, in early evolution, it is likely that RNA served this role. As a testimony for this, there still exist organisms, such as RNA viruses, that use RNA instead of DNA to carry the hereditary information from one generation to another. Retroviruses such as the Human Immunodefficiency Virus (HIV) are a subclass of RNA viruses, in which the RNA replicates via a DNA intermediate, using the enzyme **reverse transcriptase** (RT). This enzyme is an RNA-dependent DNA polymerase that, simply said, makes DNA from RNA. This enzyme plays a very important role in microarray technology as will be discussed later.

Let us re-examine the several processes that are fundamental for life and important to understand the need for microarray technology and its emergence. Every cell of an individual organism contains the same DNA, carrying the same information.[5] In fact, this is the very basis of the use of DNA as evidence in criminal cases: a single droplet of bodily fluid, a single hair, or a few cells can uniquely identify an individual. However, in spite of carrying the very same DNA, a liver cell is obviously different from a muscle cell, for example. These differences occur because not all genes are expressed in the same way in all cells. The differentiation between cells is given by different patterns of gene activations, which in turn, control the production of proteins. Much as studying the different levels of expression of various genes in different tissues can help us understand why the tissues are different, studying the different levels of expression between the same tissue in different conditions can help us understand the differences between those conditions.

Proteins are long, linear molecules that have a crucial role in all life processes. Proteins are chains of amino acid molecules. As previously discussed, there are 20 amino acid molecules that can be combined to build proteins. The number of all possible sequences of amino acids is staggering. For instance, a sequence with length 10 can contain $20^{10} = 10,240$ billion different combinations of amino acids, which is a very large number indeed. Although proteins are linear molecules, they are folded in complex ways. The protein-folding process is a crucial step since a protein has to be folded into a very specific

[5]Like many other things in life sciences, this is true *most of the time*, rather than always. There could be, in fact, individuals of various species, including humans, that can be composed of two or more populations of genetically distinct cells, that originated from different zygotes. Such an individual is called chimera, after a mythological creature that had a body composed of parts of different animals. Chimeras are extremely rare.

way for it to function properly.[6] Enzymes are specialized proteins[7] that act as catalysts and control the internal chemistry of the cells. This is usually done by binding to specific molecules in a very precise way such as certain atoms in the molecules can form bonds. Sometimes, the molecules are distorted in order to make them react more easily. Some specialized enzymes process DNA by cutting long chains into pieces or by assembling pieces into longer chains. A gene is active, or expressed, if the cell makes the gene product (e.g. protein) encoded by the gene. If a lot of gene product is produced, the gene is said to be highly expressed. If no gene product is produced, the gene is not expressed (unexpressed).

2.3.1 Transcription

The process of using the information encoded into a gene to produce a protein involves reading the DNA sequence of the gene. The first part of this process is called **transcription** and is performed by a specialized enzyme called **RNA polymerase**. Essentially, the transcription process converts the information coded into the DNA sequence of the gene into an RNA sequence. This "expression" of the gene will be determined by various internal or external factors. The objective of researchers is to detect and quantify gene expression levels under particular circumstances.

The RNA molecule is a long polynucleotide very similar to DNA, but with several important differences. First, the backbone structure of RNA is not the same. Second, RNA uses the base uracil (U) instead of thymine (T).[8] And finally, RNA molecules in cells exist as single stranded entities, in contrast to the double helix structure of DNA.

The transcription process is somewhat similar to the DNA replication. Much like in the DNA replication, the part of the DNA sequence that contains the gene to be transcribed has to be unfolded and accessible, the two DNA strands are temporarily separated and the RNA polymerase moves along one strand, reading its succession of bases and constructing an RNA sequence containing the same information. The enzyme RNA polymerase attaches itself to a specific DNA nucleotide sequence situated just before the beginning of a gene. This special sequence is called a **promoter** and it works by setting up the RNA polymerase on the correct DNA strand and pointing in the right direction. The two DNA strands are separated locally such that the RNA polymerase can do its work and transcribe DNA into RNA. The RNA molecule

[6]The so-called "mad cow disease" is apparently caused by a brain protein folded in an unusual way. When a wrongly folded protein comes into contact with a normal protein, it induces the normal protein to fold itself abnormally. This abnormal folding prevents the protein from performing its usual role in the brain, which leads to a deterioration of brain functions and, eventually, death.

[7]Most but not all enzymes are proteins. Some RNA molecules called ribozymes act like enzymes.

[8]This is why the CT tuple appears as CU in Fig. 2.11 which shows the mapping from RNA to codons, rather than DNA to codons.

is synthesized as a single strand, with the direction of transcription being $5' \rightarrow 3'$. The RNA polymerase starts constructing RNA using ribonucleotides freely available in the cell. The RNA sequence constructed will be complementary to the DNA sequence read by the RNA polymerase. When the DNA sequence contains a G for instance, the polymerase will match it with a C in the newly synthesized RNA molecule; likewise, an A will be matched with a U in the chain under construction (because U substitutes for T in the RNA molecule), etc. The process will continue with the RNA polymerase moving into the gene, reading its sequence, and constructing a complementary RNA chain until the end of the gene is reached. The end of the gene is marked by a special sequence that signals the polymerase to stop. When this sequence is encountered, the polymerase ends the synthesis of the RNA chain and detaches itself from the DNA sequence.

The RNA sequence thus constructed contains the same information as the gene. This information will be used to construct the protein coded for by the gene. However, the structure of the protein is not yet completely determined by the RNA sequence synthesized directly from the DNA sequence of the gene. The RNA chain synthesized by the RNA polymerase is called a primary transcript or pre-mRNA and is only the initial transcription product. In fact, the sequence of nucleotides in a gene may not be used in its entirety to code for the gene product. Thus, for more complex organisms, a great part of the initial RNA sequence is disposed of during a **splicing** process to yield a smaller RNA molecule called **messenger RNA** or **mRNA**. Its main role is to carry this information to some cellular structures outside the nucleus called ribosomes where proteins will be synthesized. The non-coding stretches of sequence that are eliminated from the primary transcript to form a mature mRNA are called **introns**. Conversely, the regions that will be used to build the gene product are called coding regions, or **exons**. Thus, RNA molecules transcribed from genes containing introns are longer than the mRNA that will carry the code for the construction of the protein.

The mechanism that cuts the transcribed RNA into pieces, eliminates the introns, and reassembles the exons together into mRNA is called **RNA splicing**. The RNA splicing takes place in certain places determined by a specific DNA sequence that characterizes the intron/exon boundaries. Fig. 2.14 shows the consensus sequence for the intron/exon boundaries. In general, the splicing is carried out by small, nuclear RNA particles (**snRNP**s, pronounced snurps) that get together with some proteins to form a complex called **spliceosome**. During the splicing, at each splicing site, the spliceosome bends the intron to be eliminated in the shape of a loop called lariat, bringing together the two exon ends to be connected. In a subsequent step, the two exon ends are connected, the lariat is cut off, and the spliceosome detaches from the mRNA.

Depending on the circumstances, the pre-mRNA can be cut into different pieces, and these pieces can be assembled in different ways to created different proteins. This mechanism that allows the construction of different mRNAs from the same DNA sequence is called **alternative splicing**. The mechanism

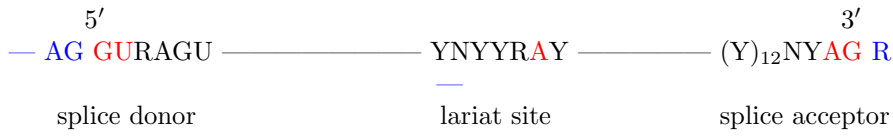


FIGURE 2.14: The consensus sequence of a splicing site. The symbol R denotes any puRine (A or G); the symbol Y denotes any pYrimidine (C or U); the symbol N stands for aNy of A, C, G, or U. The lines stand for arbitrary sequences. The blue color represents the ends of the exons. As shown in the figure, there is a lot of variability in these sites with the exception of the bases in red which are required in order for the splicing to occur. During the splicing, the spliceosome bends the intron to be eliminated in the shape of a loop called lariat, bringing together the two exon ends to be connected. In a subsequent step, the two exon ends are connected, the lariat is cut off and the spliceosome detaches from the mRNA.

of alternative splicing greatly increases the protein coding abilities of genes by allowing a gene to code for more than one protein. Fig. 2.15 shows a number of alternative **splice variants** encoded by a single gene. The pre-mRNA shown at the top of Fig. 2.15 includes a number of introns and exons. An mRNA is obtained in each case by eliminating some introns and concatenating the remaining exons. Which particular mRNA is constructed at any one time depends on the circumstances and is controlled through a number of mechanisms. Depending on their effect, these mechanisms are divided into enhancers and silencers, and subdivided by their target into exon-splicing enhancers and silencers, or intron-splicing enhancers and silencers.

Another important reaction that occurs at this stage is called **polyadenylation**. This reaction produces a long sequence of A nucleotides concatenated onto the 3′ end of a mature mRNA. This reaction is of interest because some protocols in microarray technology use the final product of polyadenylation. Transcription of the RNA is known to stop after the enzymes (and some specialized small nuclear RNAs) responsible for the transcription process recognize a specific termination site. Cleavage of the RNA molecule occurs at a site with sequence AAUAAA and then about 200 adenylate (i.e., AMP) residues are sequentially added in mammalian cells by the enzyme poly(A) polymerase to form a poly(A) tail. This tail is used as a target in the process of reverse transcription.

2.3.2 Translation

After the post-transcriptional processing, the mRNA transcribed from the genes in the nuclear DNA leaves the nucleus and moves into the cytoplasm.

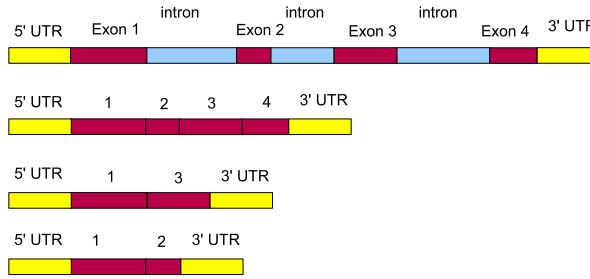


FIGURE 2.15: Alternative splice variants encoded by a single gene. The pre-mRNA shown at the top of the figure includes a number of introns and exons. The mRNA is obtained in each case by eliminating some introns and concatenating the remaining exons. In general, the splicing is carried out by small, nuclear RNA particles (snRNPs, pronounced snurps) that get together with some proteins to form a complex called spliceosome. UTRs represent untranslated regions. The mechanism of alternative splicing greatly increases the protein-coding abilities of genes by allowing a gene to code for more than one protein.

The mRNA containing the sequence coding for the protein attaches to ribosomes (see Section 2.1). Here, the information contained in the mRNA is mapped from a sequence of RNA nucleotides into a sequence of amino acids forming the protein. This process is called **translation**. As a mnemonic help, this process translates the information necessary in order to construct a protein from the 4 base alphabet of the DNA/RNA to the 20 letter alphabet of the amino acids.

The ribosome attaches to the messenger RNA near a specific start codon that signals the beginning of the coding sequence. The various amino acids that form the protein are brought to the ribosome by molecules of RNA that are specific to each type of amino acid (see Fig. 2.16). This RNA is called **transfer RNA** (**tRNA**). The tRNA molecules recognize complementary-specific codons on the mRNA and attach to the ribosome. The first tRNA to be used will have a sequence complementary to the sequence of the first codon of the mRNA. In turn, this first tRNA molecule will bring to the ribosome the first amino acid of the protein to be synthesized. Subsequently, a second tRNA molecule with a sequence complementary to the second codon on the mRNA will attach to the existing ribosome-mRNA-tRNA complex. The shape of the complex is such that the amino acids are brought into proximity and they bind to each other. Then, the first tRNA molecule is released and the first two amino acids linked to the second tRNA molecule are shifted on the ribosome bringing the third codon into position. The tRNA bringing the third amino acid can now attach to the third codon because of its complementary sequence and the process is repeated until the whole protein molecule is synthesized.

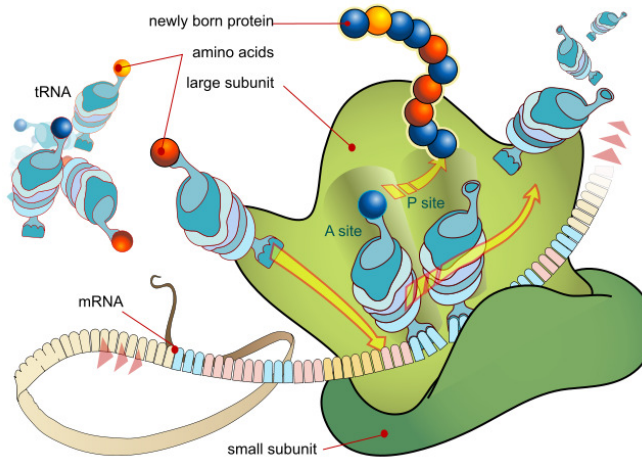


FIGURE 2.16: The protein translation process. The tRNA molecules recognize complementary-specific codons on the mRNA and attach to the ribosome. Each such tRNA molecule brings the amino acid corresponding to its respective codon, which is attached to the newly-formed polypeptide chain that will become the protein. Once the amino acid has been attached, its corresponding tRNA molecule is detached, and the process is repeated for the next amino acid, until the entire protein is assembled.

The process stops when a special stop codon is encountered, which signals the mRNA to fall off the ribosome together with the newly constructed protein.[9] After it is released, the protein may suffer a set of final changes called post-translational modifications. Such modifications might include cleavage, folding, phosphorylation, methylation, etc. Once these are done, the protein starts performing the cellular function for which it was designed. At this stage, it is said that the protein is active.

It must be mentioned that the process described above is greatly simplified. For instance, the complex between tRNA molecules and their corresponding amino acids is in turn controlled by another enzyme called aminoacyl-tRNA synthetase. There is at least one type of synthetase for each type of amino-acid. Since other complex steps, such as tRNA–amino acid reaction, are involved in the protein synthesis, it is clear that the amount of protein produced in the cell is also dependent on the successful completion of all these intermediate steps. Furthermore, as explained above, post-translational modifications can be crucial in making a protein active. Having abundant amounts of inactive protein will not help the cell perform the necessary functions. However, in

[9]See [121] for an excellent introduction to DNA and gene cloning for the nonspecialist and [406] for a more complete treatment of the subject.

general, there is a quantitative correspondence between the amount of mRNA produced by the enzyme reading the gene and the amount of protein produced. Therefore, the amount of mRNA produced from various genes is usually directly proportional to the amount of protein produced from that mRNA, i.e. to the expression level of that gene. This is the main assumption at the basis of most experiments that try to characterize the gene's expression levels using DNA microarrays. Nevertheless, one should keep in mind that the measured levels of mRNA do not always map to proportional levels of protein, and even if they did, not all those proteins may be in an active form, etc. A number of other techniques are available to obtain information at other levels of this complex process. For instance, proteomics techniques can provide information about the amounts of proteins available, phosphorylation assays can provide information about the amount of phosphorylated protein available, etc. A complete understanding of the cellular processes will inevitably require the integration of many heterogeneous types of data.

2.3.3 Gene regulation

The regulation of gene expression is the process that living cells use to control the amount of a gene product that is produced in the cell at any one time. As discussed above, most gene products are proteins. However, there are genes that produce RNA that is never translated into protein and yet they play some role in the cell. This process of controlling the amount of gene product is also called **gene modulation**. Gene regulation is continuously active for many genes in many cells. Because of gene regulation, a cell or organism can modify its response depending on environmental factors, signals from other cells or organisms, and even time of the day. For instance, *E. coli* is able to use various types of nutrients such as lactose and glucose. However, glucose is much more energy efficient so *E. coli* prefers to consume it if it's available. If glucose is not available but lactose is, *E. coli* regulates several of its genes to produce an enzyme, *β-galactosidase*, which is able to digest lactose. Similarly, the yeast can switch between a metabolism that uses oxygen to one that does not (beer versus bread).

Gene regulation can happen during any stage of the process that leads from a gene to a functional protein. During transcription, for instance, regulation can happen through one or more of the following mechanisms:

1. **Transcription factors** – these are proteins that bind DNA and control the production of RNA from DNA. These can be subdivided into:
   (a) **Repressors** – bind to the DNA strand nearby or overlapping the promoter region, preventing the RNA polymerase from transcribing the gene
   (b) **Activators** – bind to the DNA strand nearby of overlapping the promoter region facilitating the interaction between the RNA poly-

merase and a particular promoter, increasing the expression of the given gene

2. **Regulatory elements** – these are sites on the DNA helix that are involved in transcription control and regulation. These can be:
   (a) **Promoters** – - are the sites on the DNA helix that the activators bind to. Promoters are usually found close to the beginning of the gene
   (b) **Enhancers** – are also sites on the DNA helix that are bound by activators; an enhancer may be located upstream or downstream of the gene that it regulates. Furthermore, an enhancer does not need to be located near to the transcription initiation site to affect the transcription of a gene,

Gene regulation can also happen during RNA processing as well as after translation, through **post-translational modifications**. These post-translational modifications are chemical modifications that happen to the protein after it is translated. Such modifications usually involve the addition or removal of a functional group such as phosphate (phosphorylation), acetate (acetylation), lipids, carbohydrates, etc., or by making structural changes. Many proteins have an **active** form, in which they can perform their role in the cell, and an **inactive** form, in which they cannot perform their usual activity.

A more recently discovered mechanism of gene regulation involves **RNA interference** performed either through **micro-RNA**s (miRNA) [365] or **small interfering RNA**s (siRNA) [149]. The miRNAs are short (21–23 nucleotides), single-stranded, RNA molecules that are partially complementary to one or more mRNA sequences corresponding to other genes. Since the miRNA molecules will bind to their target mRNA molecules, fewer such molecules will be available for subsequent translation so the effect of the miRNAs will be to down-regulate their target genes. The siRNAs are short (20–25 base pairs), double-stranded RNA molecules that can interfere with the process of gene transcription-translation of other genes, either by degradation of the targeted RNA, or by histone and DNA methylation.

2.4 The need for high-throughput methods

Why bother measuring the expression of all genes? A simple answer involves the fact that the genomes of many model organisms have been sequenced, and we would like to simply have the luxury of looking at the whole genome expression profile under the influence of a particular factor. Several methods have long been available to measure expression levels but, alas, only for a

few genes at a time. Large-scale screenings of gene expression signatures were not possible the way they are routinely performed nowadays with microarrays. Therefore, a need for a quick snapshot of all or a large set of genes was pressing. Another important reason for the emergence of microarrays is the necessity to understand the networks of biomolecular interactions at a global scale. Each particular type of cell (e.g., tissue) will be characterized by a different pattern of gene expression levels, i.e. each type of cell will produce a different set of proteins in very specific quantities. A typical method in genetics was to use some method to render a gene inactive (knock it out) and then study the effects of this knockout in other genes and processes in a given organism. This approach, which was for a long time the only approach available, is terribly slow, expensive, and inefficient for a large-scale screening of many genes. Microarrays allow the interrogation of thousands of genes at the same time. Being able to take a snapshot of a whole gene expression pattern in a given tissue opens innumerable possibilities. One can compare various tissues with each other, or a tumor with the healthy tissue surrounding it. One can also study the effects of drugs or stressors by monitoring the gene expression levels. Gene expression can be used to understand the phenomena related to aging or fetal development. Screening tests for various conditions can be designed if those conditions are characterized by specific gene expression patterns. Drug development, diagnosis, comparative genomics, functional genomics, and many other fields may benefit enormously from a tool that allows accurate and relatively inexpensive collection of gene expression information for thousands of genes at a time.[10]

2.5 Summary

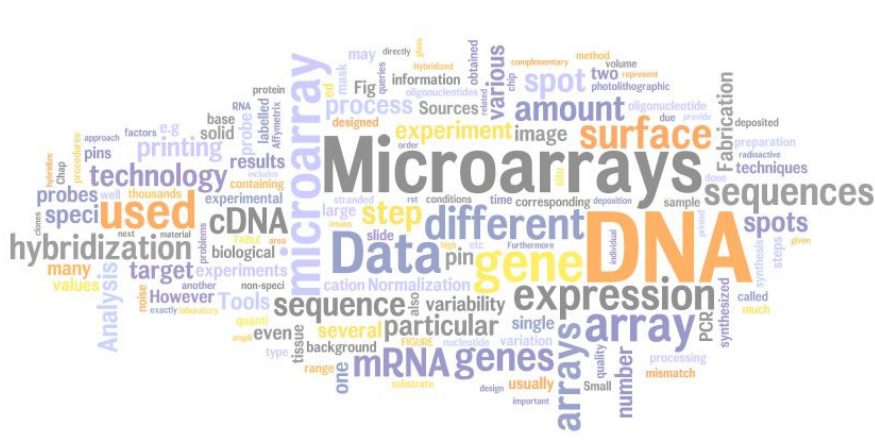
Deciphering the genomes of several organisms, including that of humans, led to an avalanche of data that needed to be analyzed and translated into biological meaning. This sparked the emergence of a new scientific field called bioinformatics. This term is generally used to denote computer methods, statistical and data mining techniques, and mathematical algorithms that are used to solve biological problems. The field of bioinformatics combines the efforts of experts from various disciplines who need to communicate with each other and understand the basic terms in their corresponding disciplines. This chapter was written for computer engineers, statisticians, and mathematicians to help them refresh their biological background knowledge. The chapter de-

[10]This is not to be interpreted that microarrays will substitute gene knockouts. Knocking out a gene allows the study of the more complex effects of the gene, well beyond the mRNA abundance level. Microarrays are invaluable as screening tools able to simultaneously interrogate thousands of genes. However, once interesting genes have been located, gene knockouts are still invaluable tools for a focused research.

scribed the basic components of a cell, the structure of DNA and RNA, and the process of gene expression. There are 4 types of DNA building blocks called nucleotide bases: A, C, G, and T. These 4 bases form the genetic alphabet. The genetic information is encoded in strings of variable length formed with letters from this alphabet. Genetic information generally flows from DNA to RNA to proteins; this is known as the central dogma of molecular biology. Genetic information is stored in various very long strings of DNA. Various substrings of such a DNA molecule constitute functional units called genes and contain information necessary to construct proteins. The process of constructing proteins from the information encoded into genes is called gene expression. First, the information is mapped from DNA to RNA. RNA is another type of molecule used to carry genetic information. Similarly to DNA, there are 4 types of RNA building blocks and a one-to-one mapping from the 4 types of DNA bases to the 4 types of RNA bases. The process of converting the genetic information contained in a gene from the DNA alphabet to the RNA alphabet is known as transcription. The result of the transcription is an RNA molecule that has the informational content of a specific gene. In higher organisms, the transcription process takes place in the cell's nucleus, where DNA resides. The RNA molecules are subsequently exported out of the cell nucleus into the cytoplasm where the information is used to construct proteins. This process, known as translation, converts the message from the 4-letter RNA alphabet to the 20-letter alphabet of the amino acids used to build proteins. The amounts of protein generated from each gene determine both the morphology and the function of a given cell. Small changes in expression levels can determine major changes at the organism level and trigger illnesses such as cancer. Therefore, comparing the expression levels of various genes between different conditions is of extreme interest to life scientists. This need stimulated the development of high throughput techniques for monitoring gene expression such as microarrays.

Chapter 3

Microarrays



If at first you don't succeed, you are running about average.

—M. H. Alderson

3.1 Microarrays – tools for gene expression analysis

In its most general form, a DNA array is usually a substrate (nylon membrane, glass or plastic) on which one deposits single-stranded DNAs (ssDNA)with various sequences. Usually, the ssDNA is printed in localized features that are arranged in a regular grid-like pattern. In this book, we will conform with the nomenclature proposed by Duggan et al. [139], and we will refer to the ssDNA printed on the solid substrate as a **probe**.

What exactly is deposited depends on the technology used and on the purpose of the array. If the purpose is to understand the way a particular set of genes function, the surface will contain a number of regions dedicated to those individual genes. However, arbitrary strands of DNA may be attached to the surface for more general queries or DNA computation. The array thus fabricated is then used to answer a specific question regarding the DNA on its surface. Usually, this interrogation is done by washing the array with a solution containing ssDNA, called a **target**, that is generated from a particular

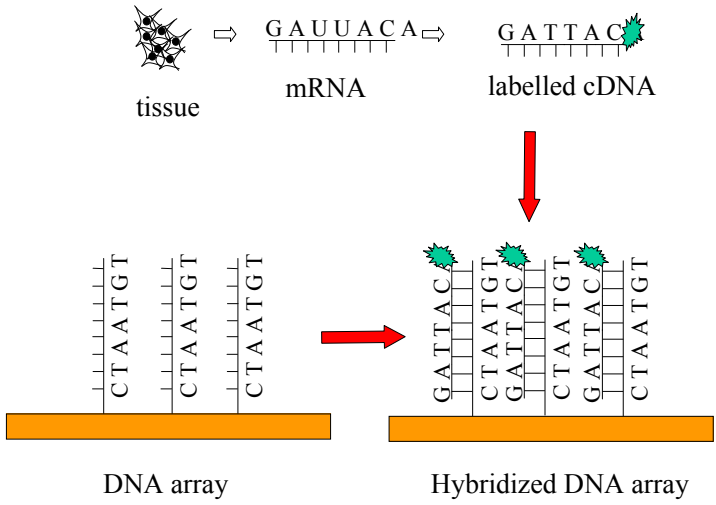


FIGURE 3.1: A general overview of the DNA array used in gene expression studies. The mRNA extracted from tissue is transformed into complementary DNA (cDNA), which is hybridized with the DNA previously spotted on the array.

biological sample under study as described below. The idea is that the DNA in the solution that contains sequences complementary to the sequences of the DNA deposited on the surface of the array will hybridize to those complementary sequences. The key to the interpretation of the microarray experiment is in the DNA material that is used to hybridize on the array. Since the target is labeled with a fluorescent dye, a radioactive element, or another method, the hybridization spot can be detected and quantified easily.

When used in gene expression studies, the DNA target used to hybridize the array is obtained by reverse transcription of the mRNA extracted from a tissue sample to a double stranded complementary DNA (cDNA) (see Fig. 3.1). This DNA is fluorescently labeled with a dye, and a subsequent illumination with an appropriate source of light will provide an image of the array of features (sets of probes on GeneChips, spots on cDNA arrays, or beads on Illumina arrays). The intensity of each spot or the average difference between matches and mismatches can be related to the amount of mRNA present in the tissue and, in turn, with the amount of protein produced by the gene corresponding to the given feature.

This step can also be accomplished in many different ways. For instance, the labeling can be done with a radioactive substance and the image obtained

by using a photosensitive device. Or several targets can be labeled with different dyes and used at the same time in a competitive hybridization process in a multichannel experiment. A typical case is a two-channel experiment using cy3 and cy5 as dyes, but other dyes can also be used. After an image-processing step is completed, the result is a large number of expression values. Typically, one DNA array will provide expression values for hundreds or thousands of genes.

3.2 Fabrication of microarrays

Two main approaches are used for microarray fabrication: deposition of DNA fragments and *in situ* synthesis. The first type of fabrication involves two methods: deposition of PCR-amplified cDNA clones, and printing of already synthesized oligonucleotides. *In situ* manufacturing can be divided into photolithography, ink-jet printing, and electrochemical synthesis.

3.2.1 Deposition

In deposition-based fabrication, the DNA is prepared away from the chip. Robots dip thin pins into the solutions containing the desired DNA material and then touch the pins onto the surface of the arrays. Small quantities of DNA are deposited on the array in the form of spots. Unlike *in situ* manufacturing in which the length of the DNA sequence is limited, spotted arrays can use small sequences, whole genes or even arbitrary PCR products.

As discussed in Chapter 2, the living organism can be divided into two large categories: eukaryotes and prokaryotes. The group of eukaryotes includes the organisms whose cells have a nucleus. Prokaryotes are organisms whose cells do not have a nucleus, such as bacteria. In general, eukaryotes have a much more complex intracellular organization than prokaryotes. Gene expression in most eukaryotes is studied by utilizing complementary DNA (cDNA) clones, which allow the amplification of sufficient quantities of DNA for deposition. Mature mRNA is reverse transcribed into short cDNAs and introduced into bacterial hosts, which are grown, isolated, then selected out if they carry foreign DNA. As discussed in Chapter 2, bacteria are prokaryotes, which, unlike eukaryotes, do not have a nucleus and do not have introns in their DNA. Therefore the prokaryotic gene expression machinery is different, and it is less complicated to amplify their genes.

The cloning strategy leverages bacterial properties in order obtain large quantities of eukaryotic DNA. Single-pass, inexpensive sequencing of entire clone libraries results in sets of **expressed sequence tags** (ESTs), which are partial sequences of the clone inserts that are long enough to uniquely identify the gene fragments. The **polymerase chain reaction** (PCR) is used to am-

plify clones containing desired fragments, using primers flanking the inserts, or oligonucleotide primers designed specifically for selective amplification. Once the cDNA cloned inserts are amplified by PCR, they are purified, and the final PCR products are then spotted on a solid support.

Another method of microarray fabrication is the attachment of short, synthesized oligonucleotides to the solid support. One advantage of this method is that oligonucleotide probes can be designed to detect multiple variant regions of a transcript or the so-called splice variants (see Fig. 2.15). These oligonucleotides are short enough to be able to target specific exons. Measuring the abundance of specific splice variants is not possible with spotted cDNA arrays because cDNA arrays contain probes of long and variable length, to which more than one different splice variant might hybridize.

3.2.1.1 The Illumina technology

A more recent technology for the fabrication of microarrays is the BeadArray technology, developed by Illumina (see Fig. 3.2). This technology uses BeadChips, which are microarrays composed of very small (3 μm) silica beads that are placed in small wells etched out in one of two substrates: fiber-optic bundles or planar silica slides. These beads are randomly self-assembled on the substrate in a uniform pattern that places the beads approximately 5.7 microns apart. Each bead is covered with hundreds of thousands of copies of a given nucleotide sequence forming the probe specific to the given assay. Each such oligonucleotide sequence is approximately 50 bp long and is concatenated with another custom-made sequence that can be used as an address. This address sequence is used to locate the position of each bead on the array and to uniquely associate each bead with a specific target site.

After the self-assembly is complete, individual bead types are decoded and identified. Figure 3.3 illustrates how the decoding process works in a simplified example using 16 different bead types. The randomly assembled array is sequentially hybridized to 16 "decoder oligonucleotides," each of which is a perfect match for one of the assay oligos bound to a particular bead type. In this example, the first four decoder oligos are labeled with the same blue fluorescent dye, and the second set of four decoder oligos are labeled with a green dye, and so on. The array is hybridized to the first set of 16 decoder oligos, labeled as described above, then imaged and stripped. The second hybridization includes the same 16 decoder oligos, labeled in a different order with fluorescent dyes. Following the second round of hybridization and imaging, it is simple to precisely identify the exact bead type in each position on the array. For example, a location that is blue in the first round and then yellow in the second round is bead type number 3, while a location that is yellow in the first round and then green in the second round is bead type number 10.

Since the space needed for each bead is so small, a high density can be achieved on the array, allowing thousands, or even millions, of target sites to

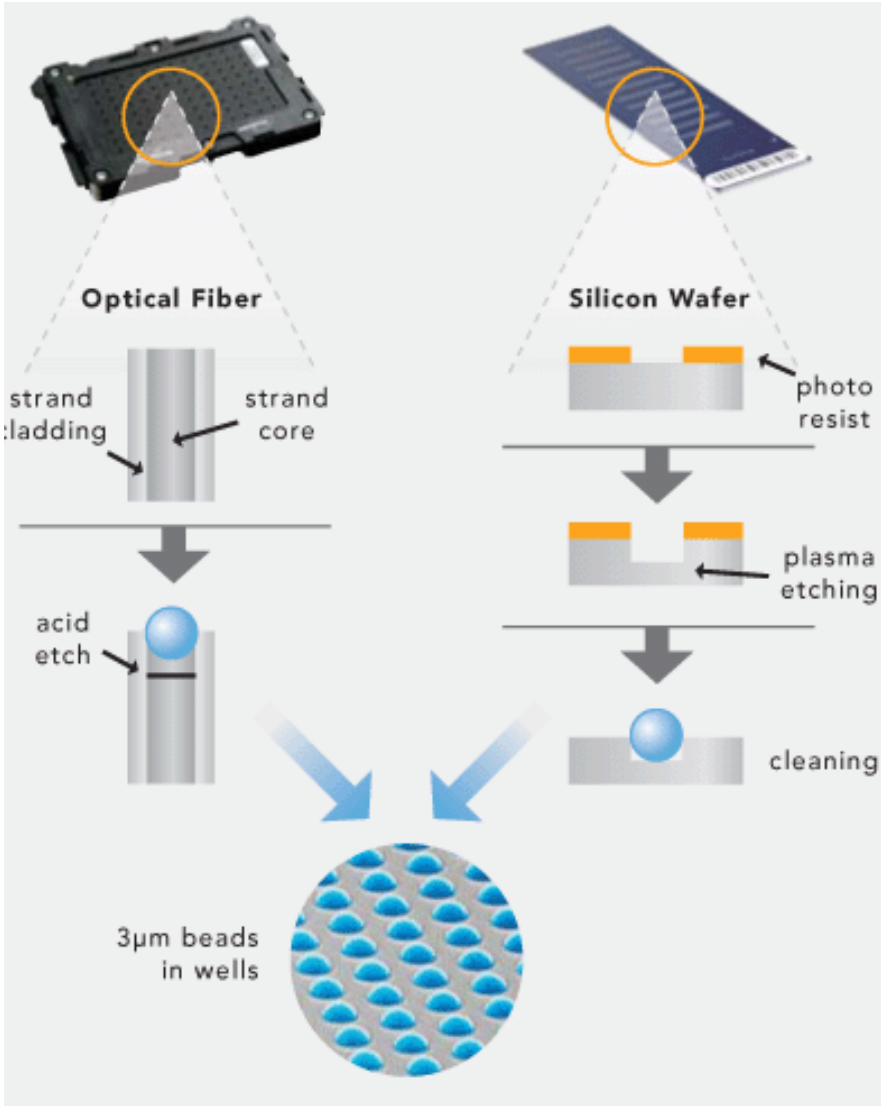


FIGURE 3.2: The Illumina BeadArray Technology. Very small (approximately 3 microns) silica beads are placed in small wells etched out of either optical fibers or a silicon wafer. The beads are held in place by Van der Waals forces as well as hydrostatic interactions with the walls of the well. The surface of each bead is covered with multiple (hundreds of thousands) copies of the sequence chosen to represent a gene. Courtesy of Illumina, Inc.

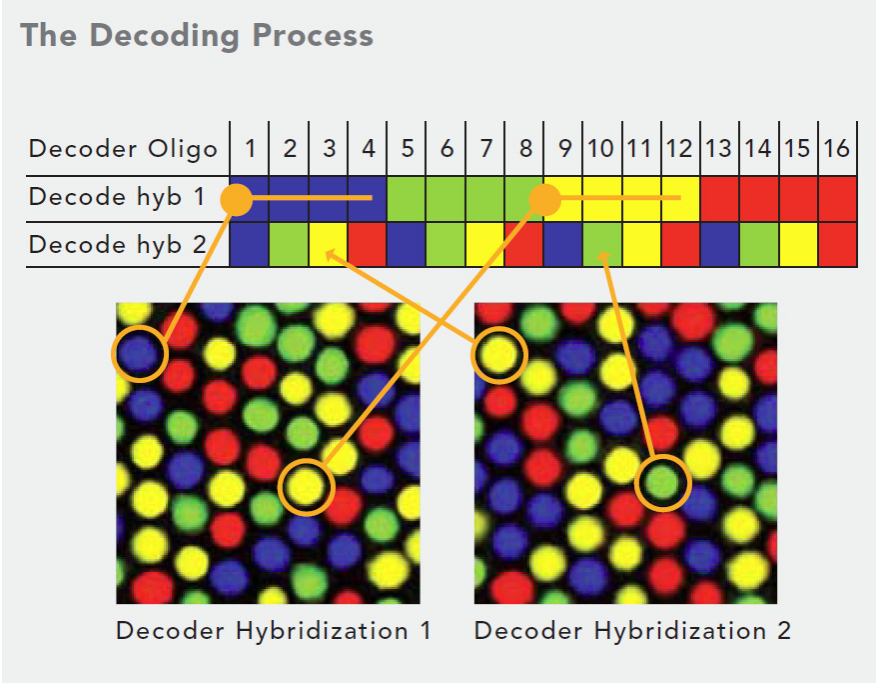


FIGURE 3.3: Decoding the Illumina bead types. The array is hybridized twice with the colors shown on the top. The combinations of colors that a given bead has in the two hybridization allows the unique identification of their type. For instance, the bead near the top left corner was blue in the first hybridization and yellow in the second one. The only type that matches this is 3. The second bead shown was yellow in the first hybridization and green in the second one yielding type 10. Courtesy of Illumina, Inc.

FIGURE 3.4: An Illumina Direct Hybridization probe is more similar to the probes used in other manufacturing technologies with the difference that the probe is attached to a bead, rather than to a flat surface. The beads are either fixed in optical fibers or distributed across a silicon wafer as shown in Fig. 3.2. The labeled target cDNA reverse-transcribed from the sample RNA hybridizes to the probe as usual. Courtesy of Illumina, Inc.

be analyzed simultaneously. Furthermore, the arrays can be formatted to test several samples in parallel.

Illumina produces whole-genome arrays, as well as more focused arrays that include probes for a subset of genes related to the specific condition. There are either one or two probes per gene, depending on the type of array (focused set or whole genome). For gene expression analysis, the Illumina arrays come in two flavors, using slightly different approaches. The Direct Hybridization Assay, illustrated in Fig. 3.4, uses a single DNA sequence per bead much like in all other technologies. This single-stranded sequence is meant to hybridize with the labeled target sequence present in the sample. The amount of fluorescence produced will provide a measure of the amount of target present in the sample.

The other approach to expression level measurements is called DASL, which stands for c**D**NA-mediated **A**nnealing, **S**election, **E**xtension and **L**igation [153]. This approach is described in Fig. 3.5 and Fig. 3.6. In the Whole-Genome DASL HT (WG DASL) Assay,[1] a pair of oligonucleotides is annealed to each target site, and more than 29,000 oligonucleotide pairs can be multiplexed together in a single reaction. A high specificity is obtained by requiring that both members of an oligonucleotide pair must hybridize in close proximity for the assay to generate a strong signal. The main advantage of the DASL approach with respect to other commercially available assays is related to the quality of the mRNA that can be evaluated. Since the WG DASL Assay uses two short sequences that in the gene are separated by a gap, there is a lot of flexibility in choosing the sequences. Furthermore, since these probes span only about 50 bases, partially degraded RNA, such as that from formalin-fixed paraffin-embedded (FFPE) samples, can be used in the

[1]More information about the WG DASL can be found at: `http://www.illumina.com/documents/products/datasheets/datasheet_whole_genome_dasl_ht.pdf`

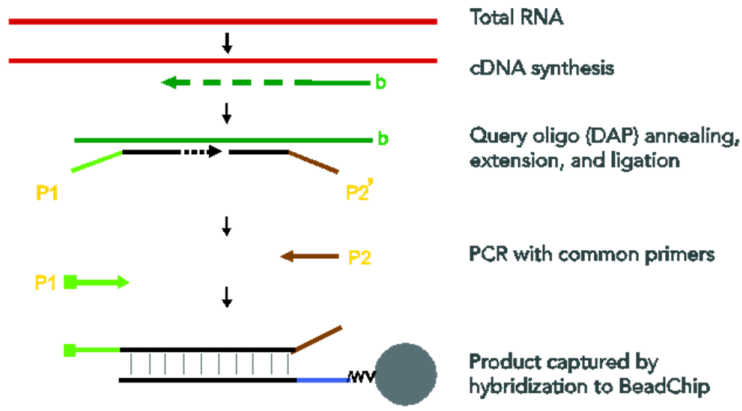


FIGURE 3.5: The Illumina WT DASL Technology. A pair of oligonucleotides is annealed to each target site. A high specificity is obtained by requiring that both members of an oligonucleotide pair must hybridize in close proximity for the assay to generate a strong signal. Since these probes span only about 50 bases, partially degraded RNA, such as that from formalin-fixed paraffin-embedded (FFPE) samples, can be used in the assay. Courtesy of Illumina, Inc.

assay. There are estimated to be more than 400 million of these FFPE samples archived in North America for cancer alone. Many of these samples represent clinical outcomes with the potential to provide critical insight into expression profiles associated with complex disease development. Unfortunately, FFPE archival methods often lead to partial RNA degradation, often limiting the amount of information that can be derived from such samples. In contrast to the WG DASL arrays, cDNA arrays using very long sequences require good quality RNA that can only be obtained from fresh tissue, or tissue frozen very soon after collection.

Genotyping with Illumina Arrays. In addition to gene expression analysis, the Illumina BeadArray platform can also be used for genotyping applications as well. The genotyping arrays span a much larger multiplex range than the expression arrays (up to 5 million markers per sample). Illumina offers two genotyping assays with the BeadArray platform: the GoldenGate Assay for custom, low-multiplex studies, and the Infinium HD Assay for high-multiplex studies.

The GoldenGate Assay. The Illumina GoldenGate Genotyping Assay is a flexible, pre-optimized assay that uses a discriminatory DNA polymerase and ligase to interrogate up to 3,072 SNP loci simultaneously (see Section 2.2.4 for more details about SNPs). This assay is illustrated in Fig. 3.7. The genomic

FIGURE 3.6: An Illumina DASL probe. The address is a short sequence that is specific to each targeted site. The address is used to attach the probes to the beads. In the DASL assay, each targeted site is represented by two sequences, one from upstream of the targeted site and one from downstream of it. Because each targeted site is represented by two short sequences that can be separated by an arbitrary gap, this assay is better able to work with partially degraded RNA, such as the one coming from older formalin-fixed paraffin-embedded samples. In contrast, cDNA arrays using long sequences usually require good-quality mRNA. Courtesy of Illumina, Inc.

DNA (gDNA) sample used in this assay is first fragmented and then bound to paramagnetic particles in preparation for hybridization with the assay oligonucleotides. Three oligonucleotides are designed for each SNP locus. Two oligos are specific to each allele of the SNP site, called the Allele-Specific Oligos (ASOs). A third oligo that hybridizes several bases downstream from the SNP site is the Locus-Specific Oligo (LSO). All three oligonucleotide sequences contain regions of genomic complementarity and universal PCR primer sites; the LSO also contains a unique address sequence that targets a particular bead type on the array. Up to 3,072 SNPs may be interrogated simultaneously in this manner using GoldenGate technology. During the hybridization process, the assay oligonucleotides hybridize to the genomic DNA sample bound to paramagnetic particles. Because hybridization occurs prior to any amplification steps, no amplification bias can be introduced into the assay. Following hybridization, extension of the appropriate ASO (the one containing the complementary SNP) and ligation of the extended product to the LSO joins information about the genotype present at the SNP site to the address sequence on the LSO. These joined, full-length products provide a template for PCR using universal PCR primers P1, P2, and P3.Universal PCR primers P1 and P2 are Cy3- and Cy5-labeled. After downstream-processing, the single-stranded, dye-labeled DNAs are hybridized to their complement bead type through their unique address sequences. Hybridization of the GoldenGate Assay products onto the BeadChip allows for the separation of the assay products in solution, onto a solid surface for individual SNP genotype readout. After hybridization, a high-precision scanner is used to analyze fluorescence signal on the BeadChip, which is in turn analyzed using software for automated genotype clustering and calling. The GoldenGate assay is designed for low-plex, custom

studies. It should be mentioned that the DASL assay described above actually uses the GoldenGate extension/ligation chemistry.

The Infinium HD Assay. The Infinium HD assay is designed for high-multiplex studies, with the ability to assay up to 5 million markers simultaneously. This assay is illustrated in Fig. 3.8.

Genomic markers are interrogated though a two-step detection process. Carefully designed 50-mer probes selectively hybridize to the loci of interest, stopping one base before the interrogated marker. Marker specificity is conferred by enzymatic single-base extension to incorporate a labeled nucleotide. Subsequent dual-color florescent staining allows the labeled nucleotide to be detected by Illumina's imaging scanners, which identify both color and signal intensity. For genotyping assays, the red and green color signals specify each allele, where homozygotes are indicated by red/red or green/green signals, and heterozyotes are indicated by red/green (yellow) signals. Signal intensity information can be used to detect structural aberrations, such as copy number variants, inversions, or translocations.

3.2.2 *In situ* synthesis

During array fabrication based on *in situ* synthesis, the probes are photochemically synthesized on the chip. There is no cloning, no spotting, and no PCR carried out, which is advantageous since these steps introduce a lot of noise in the cDNA system.

Probe selection is performed based on sequence information alone. This means that every probe synthesized on the array is known in contrast to cDNA arrays, which deal with expressed sequence tags, and, in many cases, the function of the sequence corresponding to a spot is unknown. Additionally, this technology can distinguish and quantitatively monitor closely related genes just because it can avoid identical sequence among gene family members.

There are currently three approaches to *in situ* probe synthesis. The first method is photolithographic (Affymetrix, Santa Clara, CA) and is similar to the technology used to build very large scale integrated (VLSI) circuits used in modern computers. This fabrication process uses photolithographic masks for each base. If a probe should have a given base, the corresponding mask will have a hole allowing the base to be deposited at that location. Subsequent masks will construct the sequences base by base. This technology allows the fabrication of very high density arrays but the length of the DNA sequence constructed is limited. This is because the probability of introducing an error at each step, while very small, is different from zero. In order to limit the overall probability of an error, one needs to limit the length of the sequences. To compensate for this, a gene is represented by several such short sequences. The particular sequences must be chosen carefully to avoid cross-hybridization between genes.

The second approach is the ink-jet technology (Agilent, Protogene, etc.), which employs the technology used in ink-jet color printers. Four cartridges

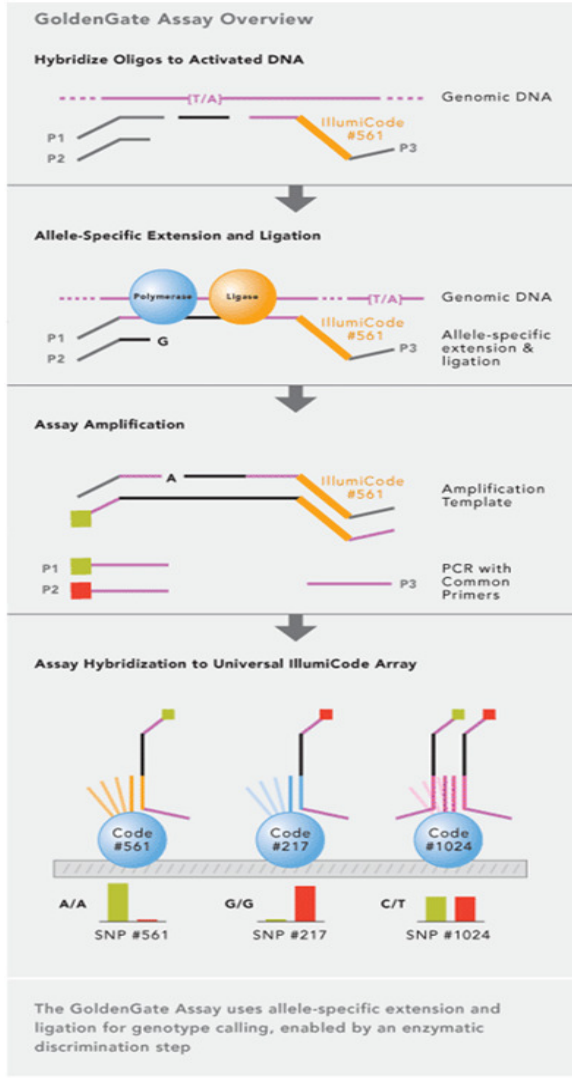


FIGURE 3.7: The Illumina Golden Gate assay. The GoldenGate assay is based on the BeadArray technology: assay oligonucleotides, containing an address sequence, hybridize to gDNA to identify the allele at a given loci. After processing and amplification, the amplified product binds to a bead on the array that contains a complementary address sequence. Dual-color fluorescence dyes, specific to each ASO, indicate the genotype of the SNP from the gDNA fragment. Courtesy of Illumina, Inc.

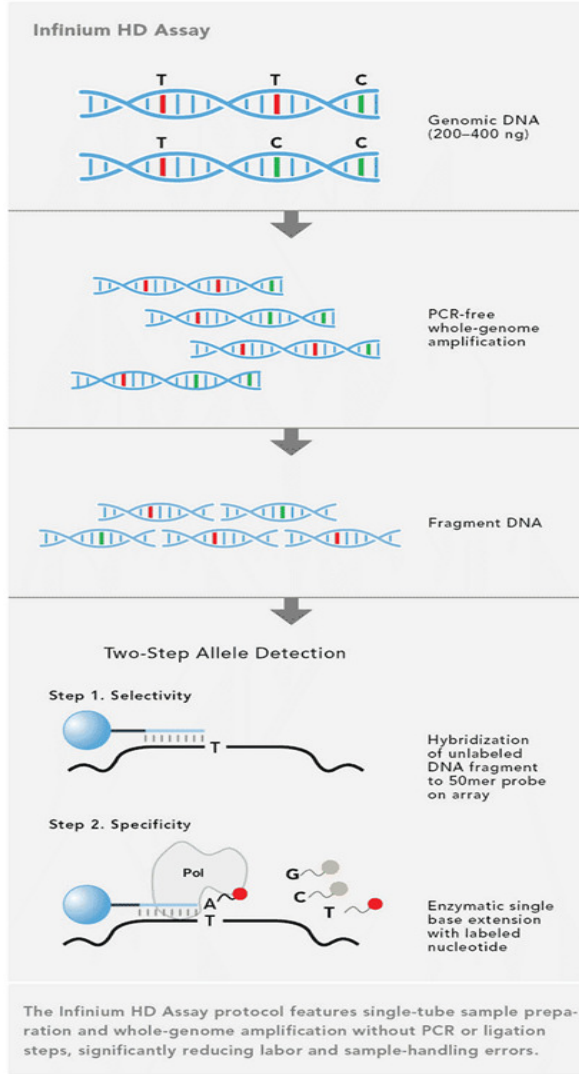


FIGURE 3.8: The Illumina Infinium assay. Genomic markers are interrogated though a two-step detection process. 50-mer probes selectively hybridize to the loci of interest, stopping one base before the interrogated marker. Subsequent dual-color fluorescent staining allows the labeled nucleotide to be detected by Illumina's imaging scanners, which identify both color and signal intensity. Courtesy of Illumina, Inc.

are loaded with different nucleotides (A, C, G, and T). As the print head moves across the array substrate, specific nucleotides are deposited where they are needed.

Finally, the electrochemical synthesis approach (CombiMatrix, Bothel, WA) uses small electrodes embedded into the substrate to manage individual reaction sites. Solutions containing specific bases are washed over the surface and the electrodes are activated in the necessary positions in a predetermined sequence that allows the sequences to be constructed base by base.

The Affymetrix technology includes the steps outlined in Figs. 3.9, 3.10, and 3.11. Synthetic linkers modified with photochemical removable protecting groups are attached to a glass surface. Light is shed through a photolithographic mask to a specific area on the surface to produce a localized photodeprotection (Fig. 3.9). The first of a series of hydroxyl-protected deoxynucleosides is incubated on the surface. In this example, it is the protected deoxynucleoside T. In the next step, the mask is directed to another region of the substrate by a new mask, and the chemical cycle is repeated (Fig. 3.10). Thus, one nucleotide after another is added until the desired chain is synthesized. Recall that the sequence of this nucleotide corresponds to a part of a gene in the organism under scientific investigation. The synthesized oligonucleotides are called probes. The material that is hybridized to the array (the reverse transcribed mRNA) is called the target, or the sample.

The **gene expression arrays** have a match/mismatch probe strategy. This is illustrated in Fig. 3.11. Probes that match the target sequence exactly are referred to as reference probes. For each reference probe, there is a probe containing a nucleotide change at the central base position – such a probe is called a mismatch. These two probes – reference and mismatch – are always synthesized adjacent to each other to control for spatial differences in hybridization. Additionally, the presence of several such pairs per gene (each pair corresponding to various parts – or exons – of the gene) helps to enhance the confidence in detection of the specific signal from background in case of weak signals.

More recently, technological advances allowed the fabrication of oligonucleotide arrays with extremely large numbers of features. At the same time, the sequencing of the entire genome has been completed for several organisms of interest, including *Homo sapiens*. Given both facts above, one could envisage arrays that cover the entire genome of a given organism. In fact, Affymetrix currently manufactures and sells such arrays, called **tiling arrays**. Affymetrix tiling arrays use short sequences, currently 25-mer oligonucleotides, that are equally spaced across the entire genome (see Fig. 3.12). The gap between two such sequences in the genome is referred to as the **resolution** of the tiling array. At the moment of this writing, Affymetrix offers tiling arrays with a resolution of 35 base pairs.

Unlike the expression arrays where probes are designed considering the direction of the strand each gene is on, tiling arrays are designed based on the direction of the genome, rather than that of a particular transcript. Tiling

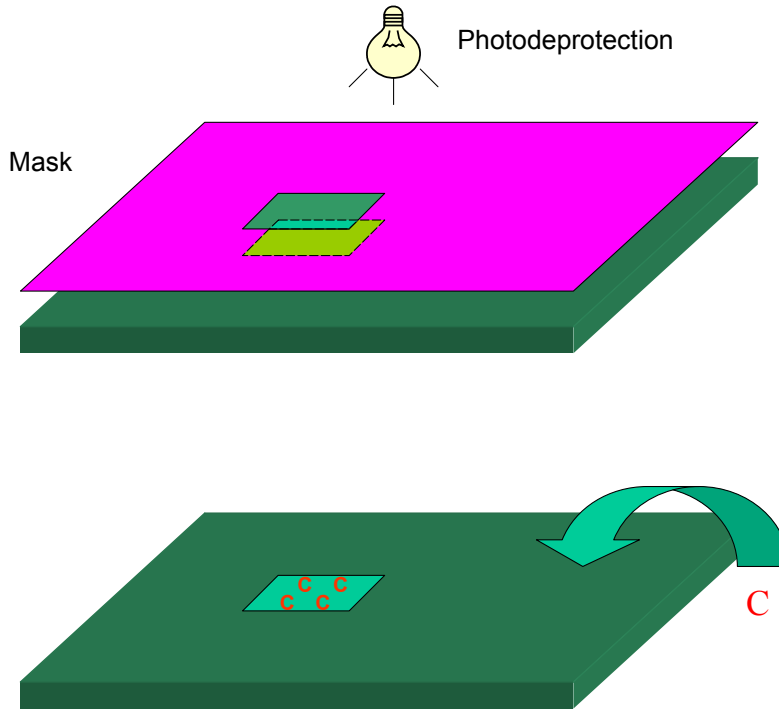


FIGURE 3.9: Photolithographic fabrication of microarrays. Synthetic linkers modified with photochemical removable protecting groups are attached to a glass surface. Light is shed through a photolithographic mask to a specific area on the surface to produce a localized photodeprotection. The first of a series of hydroxyl-protected deoxynucleosides is incubated on the surface. In this example, it is the protected deoxynucleoside C. The surface of the array is protected again, and the array is ready for the next mask.

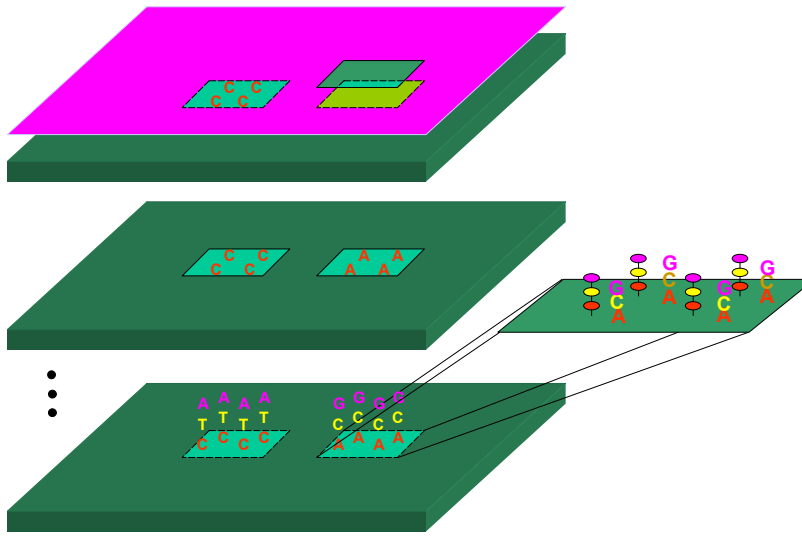


FIGURE 3.10: Photolithographic fabrications of microarrays. The second mask is applied and light is used to deprotect the areas that are designed to receive the next nucleoside (A). The fabrication process would generally require 4 masking steps for each element of the probes. Several steps later, each area has its own sequence as designed.

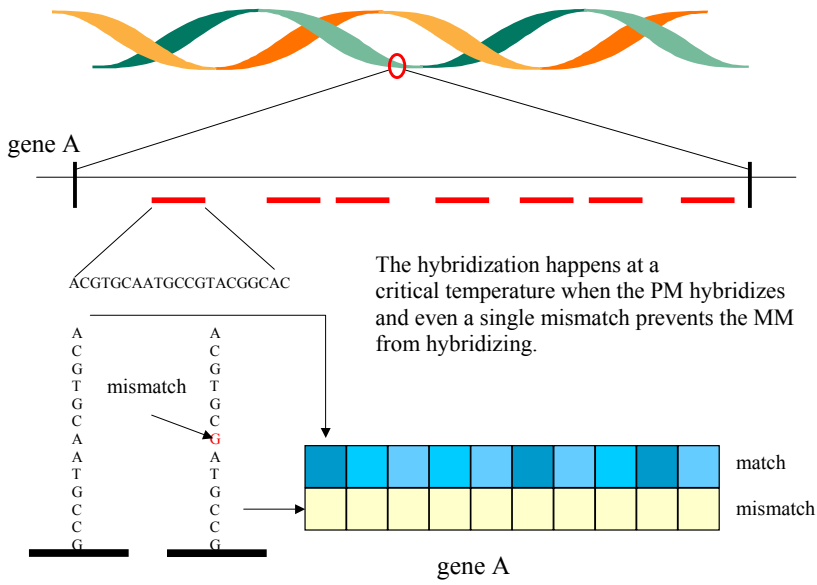


FIGURE 3.11: The principles of the Affymetrix technology. The probes correspond to short oligonucleotide sequences thought to be representative for the given gene. Each oligonucleotide sequence is represented by two probes: one with the exact sequence of the chosen fragment of the gene (perfect match or PM) and one with a mismatch nucleotide in the middle of the fragment (mismatch or MM). For each gene, the value that is usually taken as representative for the expression level of the gene is the average difference between PM and MM. Reprinted from S. Draghici, "Statistical intelligence: effective analysis of high-density microarray data" published in *Drug Discovery Today*, Vol. 7, No. 11, p. S55–S63, 2002, with permission from Elsevier.

cDNA arrays	Oligonucleotide arrays
Long sequences	Short sequences due to the limitations of the synthesis technology
Spot unknown sequences	Spot known sequences
More variability in the system	More reliable data
Easier to analyze with appropriate experimental design	More difficult to analyze

TABLE 3.1: A comparison between cDNA and oligonucleotides arrays.

arrays labeled with "F" are complementary to the forward direction (+), while tiling arrays labeled with "R" are complementary to the reverse (−) strand of the given genome. At the time of this writing, there are several tiling arrays available. The GeneChip Human Tiling 1.0R Array Set is a set of 14 arrays that include both a perfect match as well as a mismatch probe for each target location on the genome. These arrays can be used for both transcript mapping as well as in **chromatin immunoprecipitation** (ChIP) experiments. The GeneChip Human Tiling 2.0R Array Set is a set of 7 arrays that include only the perfect match probes for each location. Both sets cover the genomic sequence left after the repetitive elements were removed by RepeatMasker. Each array within the sets above contain more than 6.5 million probes. Another tiling array, the GeneChip Human Promoter 1.0R, uses the same tiling technique but focuses only on the known human promoter regions. This array includes approximatively 4.6 million probes covering 22,500 known promoter regions.

3.2.3 A brief comparison of cDNA and oligonucleotide technologies

It is difficult to make a judgment as to the superiority of a given technology. At this point in time, the cDNA technology seems to be more flexible, allowing spotting of almost any PCR product whereas the Affymetrix technology seems more reliable and easier to use. This field is so dynamic that this situation might change rapidly in the near future. Table 3.1 summarizes the advantages and disadvantages of cDNA and high-density oligonucleotide arrays. Table 3.2 shows the current performance of the Affymetrix oligonucleotides arrays [36, 286].

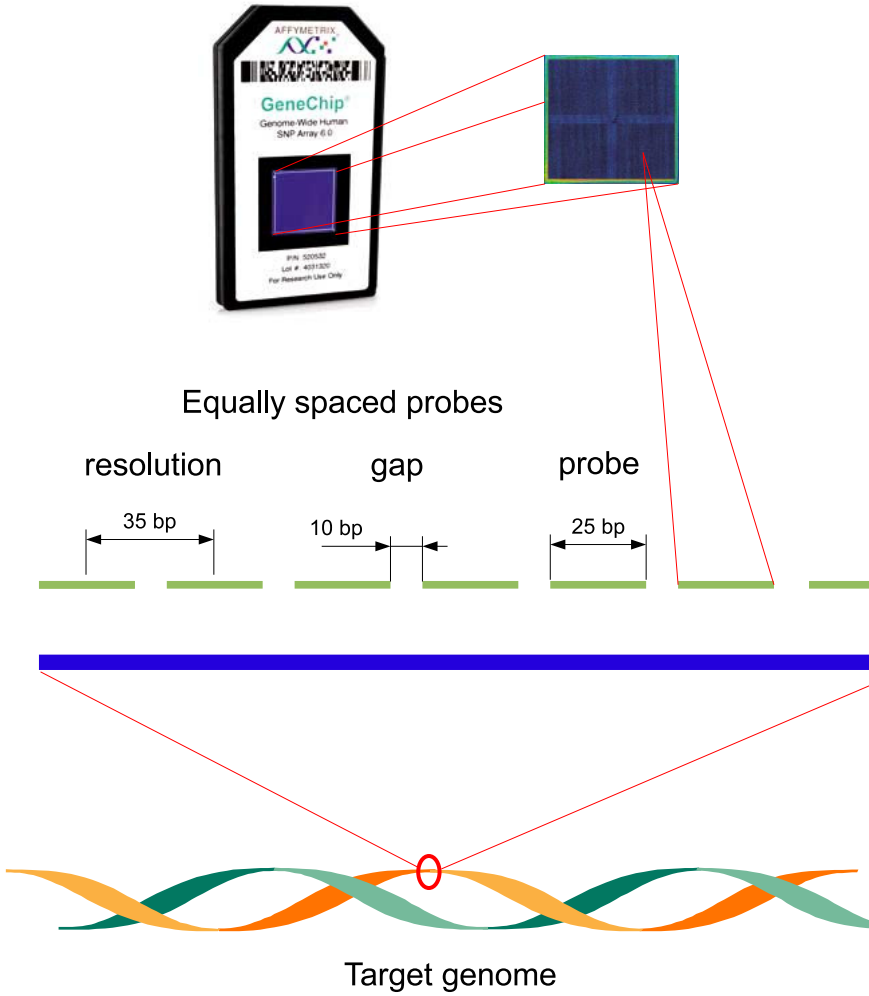


FIGURE 3.12: Tiling arrays cover the entire length of the genome after the repetitive elements have been removed. Probes of 25 oligonucleotides are tiled at an average resolution of 35 bps, with an average gap of 10 bps. Some tiling array sets contain both perfect match and mismatch sequences. Others contain only perfect matches.

	Current limit	Practical use
Density (genes/array)	40,000	20,000
Gene representation (probe pairs/gene)	4	2
Linear dynamic range	4 logs	3 logs
Fold change detection	10%	100%
Similar sequences separation	93% identical	70–80% identical
Starting material	$2ng$ total RNA	$5\mu g$ total RNA
Detection specificity	$1:10^6$	$1:10^5$

TABLE 3.2: The performance of the Affymetrix technology.

3.3 Applications of microarrays

Microarrays have been used successfully in a range of applications, including sequencing [373], SNP detection [152, 443], genotyping, disease association [386, 463], genetic linkage, genomic loss and amplification (copy number variation (CNV)) [388, 442, 449], detection of chromosomal rearrangements, etc. However, this book will focus on the (arguably) mainstream application for microarrays, which is the investigation of the genetic mechanisms in the living cells through expression analysis [147, 289, 374, 373, 384, 183, 413]. A few typical examples would include comparing healthy and malignant tissue [14, 183, 16, 54, 339], studying cell phenomena over time [113, 400] as well as study the effect of various factors such as interferons [112], cytomegalovirus infection [488], and oncogene transfection [260] on the overall pattern of expression. Perhaps even more important than the success in any individual application, the large number of papers reporting results obtained with microarrays (and subsequently validated) have increased the overall confidence that microarrays are tools that can be used to generate accurate, precise, and reliable gene expression data [84, 479, 373, 392, 391].

Microarrays can also be used for purely computational purposes such as in the field of DNA computing [249]. In these cases, the microarray can contain sequences of DNA encoding various possible solutions of the problem to be solved. Several successive steps are performed in order to solve the problem. Each such step consists of three sub-steps: a hybridization, the destruction of the single-stranded DNA not hybridized, and a denaturation that will prepare the chip for the next computational step. The role of the DNA used in each step is to prune the large number of potential solutions coded on the surface of the array. Specific sequences added in a specific step hybridize to the single-stranded DNA attached to the surface. This marks the partial solutions by binding them in double strands. Subsequently, the chip is washed with a solution that destroys the single-stranded DNA. A denaturation step will break the double-stranded DNA and bring the chip to a state in which it is ready for the next computational step.

In this book, we will concentrate on the use of microarrays in gene expression studies, focusing on specific challenges that are related to this particular application. Although the microarray data will be our main motivation and source of examples, the concepts discussed in this book, as well all analysis methods presented, are general and can be applied to a very large class of data.

3.4 Challenges in using microarrays in gene expression studies

Compared to other molecular biology techniques, microarrays are relatively new. As such, their users are challenged by a number of issues as follows:

1. **Noise.**

   Because of their nature, microarrays tend to be very noisy. Even if an experiment is performed twice with exactly the same materials and preparations in exactly the same conditions, it is likely that after the scanning and image processing steps, many genes will probably be characterized by different quantification values. In reality, noise is introduced at each step of various procedures[2] [377]: mRNA preparation (tissues, kits, and procedures vary), transcription (inherent variation in the reaction, enzymes), labeling (type and age of label), amplification, pin type (quill, ring, ink-jet), surface chemistry, humidity, target volume (fluctuates even for the same pin), slide inhomogeneities (slide production), target fixation, hybridization parameters (time, temperature, buffering, etc.), unspecific hybridization (labeled cDNA hybridized on areas that do not contain perfectly complementary sequences), nonspecific background hybridization (e.g., bleeding with radioactive materials), artifacts (dust), scanning (gain settings, dynamic range limitations, inter-channel alignment), segmentation (feature/background separation), quantification (mean, median, percentile of the pixels in one spot), etc.

   The challenge appears when comparing different tissues or different experiments. Is the variation of a particular gene due to the noise or is it a genuine difference between the different conditions tested? Furthermore, when looking at a specific gene, how much of the measured variance is due to the gene regulation and how much to noise? The noise is an inescapable phenomenon and the only weapon that the researcher seems to have against it is replication (Chapters 13 and 21).

[2]Not all steps apply to all types of arrays.

2. **Normalization.**

   The aim of the normalization is to account for systematic differences across different data sets (e.g. overall intensity) and eliminate artifacts (e.g., nonlinear dye effects). The normalization is crucial if results of different experimental techniques are to be combined. While everybody agrees on the goal of normalization, the consensus seems to disappear regarding how exactly the normalization should be done. Normalization can be necessary for different reasons such as different quantities of mRNA (leading to different mean intensities), dye nonlinearity and saturation towards the extremities of the range, etc. Normalization issues and procedures are discussed in detail in Chapter 20.

3. **Experimental design.**

   The experimental design is a crucial but often neglected phase in microarray experiments. A designed experiment is a test or several tests in which a researcher makes purposeful changes to the input variables of a process or a system in order to observe and identify the reasons for changes in the output response. Experiment design issues are discussed in details in Chapter 15.

4. **Large number of genes.**

   The fact that microarrays can interrogate thousands of genes in parallel is one of the features that led to the wide adoption of this technology. However, this characteristic is also a challenge. The classical metaphor of the needle in the haystack can easily become an accurate description of the task at hand when tens of thousands of genes are investigated. Furthermore, the sheer number of genes can change the quality of the phenomenon and the methods that need to be used. The classical example is that of the p-values in a multiple testing situation (Chapter 16).

5. **Significance.**

   If microarrays are used to characterize specific conditions (e.g., [14, 183]), a crucial question is whether the expression profiles differ in a significant way between the groups considered. The classical statistical techniques that were designed to answer such questions (e.g., chi-square tests) cannot be applied directly because in microarray experiments the number of variables (usually thousands of genes) is much greater than the number of experiments (usually tens of experiments). Novel techniques need to be developed in order to address such problems.

6. **Biological factors.**

   In spite of their many advantages, microarrays are not necessarily able to completely substitute other tools in the arsenal of the molecular biologist. For instance, knocking out genes is slow and expensive but offers an unparalleled way of studying the effects of a gene well beyond its mRNA