
Preface

Recently, the Human Genome Project announced one of the most stunning achievements in the history of science: the generation of a reference sequence of the human genome. This has brought unprecedented interest to the area of genomics, which concerns the study of large sets of genes with the goal of understanding collective function, rather than that of individual genes. Such a study is important since cellular control and its failure in disease result from multivariate activity among cohorts of genes. Very recent research indicates that engineering approaches for prediction, signal processing and control are quite well suited for studying this kind of multivariate interaction. The aim of this book is to provide the readers with an introduction to genomic signal processing, including a state-of-the-art account of the use of control theory to obtain optimal intervention strategies for gene regulatory networks, and to make readers aware of some of the open-research challenges.

Successful realization of the full potential of the area of genomics will require the collective skill and creativity of a diverse set of researchers such as biologists (including medical practitioners), statisticians and engineers. This can be possible only when each of these groups is willing and able to venture out of its immediate area of expertise, a task that is not always easy to undertake. We consider ourselves to be mathematically trained engineers who have had the opportunity to learn some basic biology while pursuing engineering research motivated by medical applications. Our main motivation for writing this book is to enable other interested readers with a similar mathematical background to get a reasonably good working knowledge of the genomics-related engineering research while having to expend only a small fraction of the time and effort that we ourselves had to originally invest.

Since the book is targeted primarily at readers whose biology background is practically non-existent, we felt that the book could be made self-contained only by including a substantial amount of material on molecular biology. However, it should be pointed out at the very outset that our molecular biology presentation, which is to a substantial extent gleaned from the classic, *Essential Cell Biology*, by Alberts et. al. [1], is *molecular biology seen through the eyes of engineers* and is the minimum needed for appropriately motivating the genomics-related engineering research presented here. As such, it is our sincere hope that the serious reader, after gaining the skeletal molecular biology knowledge from this text, will be motivated to consult other books such as the one by Alberts et. al. for additional details.

The book provides a tutorial introduction to the current engineering re-

search in genomics. The necessary molecular biology background is presented, and techniques from signal processing and control are used to (i) unearth inter-gene relationships, (ii) carry out gene based classification of disease, (iii) model genetic regulatory networks and (iv) alter (i.e. control) their dynamic behavior. The book can be divided into two parts. In the first eleven chapters, the focus is on building up the necessary molecular biology background. No prior exposure to molecular biology is assumed. In the last six chapters, the focus is on discussing the application of engineering approaches for attacking some of the challenging research problems that arise in genomics-related research.

The book begins with a basic review of organic chemistry leading to the introduction of DNA, RNA and proteins. This is followed by a description of the processes of transcription and translation and the genetic code that is used to carry out the latter. Control of gene expression is also discussed. Genetic engineering tools such as microarrays, PCR, etc., are introduced and cell cycle control and tissue renewal in multi-cellular organisms is discussed. Cancer is introduced as the breakdown of normal cell cycle control. Having covered the basics of genomics, the book proceeds to engineering applications.

The engineering techniques of classification and clustering are shown to be appropriate for carrying out gene-based disease classification. Classification then leads naturally to expression prediction which in turn leads to genetic regulatory networks. Finally, the book concludes with a discussion of control approaches that can be used to alter the behavior of such networks in the hope that this alteration will move the network from undesirable (diseased) states to more desirable (disease-free) ones.

Several people contributed to the writing of this book in many different ways. First and foremost, we would like to thank the authors of *Essential Cell Biology* and Garland Publishers, Inc. for allowing us to use a large number of figures from that book. These figures, which have been identified by proper citations at appropriate places in the text, are so informative that it is hard to imagine how we could have effectively described some of the related material without being able to refer to them. Several of our doctoral students helped at different stages with the preparation of the book. Yufei Xiao helped during the initial stages with the preparation of the first set of class notes that ultimately became this book, Ashish Choudhary was a major player in helping us get the book to its final form, and Ranadip Pal assisted us during the intermediate stages on an as-needed basis.

We would like to thank R. Kishan Baheti, Director of the Power, Control and Adaptive Networks Program and John Cozzens, Director of the Theoretical Foundations Cluster Program at the National Science Foundation for supporting our research (Grant Nos. ECS-0355227 and CCF-0514644). Financial support from the National Cancer Institute (Grant No. CA90301), the National Human Genome Research Institute, the Translational Genomics Research Institute, the University of Texas M. D. Anderson Cancer Center, and the Texas A&M Department of Electrical and Computer Engineering is

also thankfully acknowledged. Last, but not the least, we would like to thank our families without whose understanding and patience, this project could not have been completed.

Aniruddha Datta and Edward R. Dougherty

College Station, Texas

June 2006.

1. Introduction	1
2. Review of Organic Chemistry	5
2.1. Electrophilic and Nucleophilic Bonds	6
2.2. Some Common Bonds and Groups Commonly Encountered in Biological Molecules	9
2.3. Building Blocks for Common Organic Molecules	12
2.3.1. Sugars	13
2.3.2. Fatty Acids	17
2.3.3. Amino Acids	18
2.3.4. Nucleotides	20
3. Energy Considerations in Biochemical Reactions	27
3.1. Gibbs Free Energy: Thermodynamic Relevance	28
3.1.1. Phase Equilibrium	28
3.1.2. Cellular Respiration	31
3.1.3. Coupling and Reactions	32
3.2. Rate of Reactions	36
3.3. Feasibility of Chemical Reactions	37
3.4. Activated Complexes, Transition States and Their Role in Biochemistry	39
4. Proteins	37
4.1. Protein Synthesis and Function	37
4.1.1. The Amino Acid Side Chain	38
4.2. Levels of Organization in Proteins	41
4.3. Protein-Protein Interactions	44
4.4. Isolating Proteins from Cells	45
4.5. Separating a Mixture of Proteins	46
4.5.1. Gel Electrophoresis	47
4.5.2. Gel Electrophoresis	47
4.6. Protein Structure Determination	48
4.7. Proteins That Are Enzymes	49
5. DNA	53