Preface

Transcriptional regulation controls the basic processes of life. Its complex, dynamic, and hierarchical networks control the momentary availability of messenger RNAs for protein synthesis. Transcriptional regulation is key to cell division, development, tissue differentiation, and cancer as discussed in **Chapters 1** and **2**.

We have witnessed rapid, major developments at the intersection of computational biology, experimental technology, and statistics. A decade ago, researches were struggling with notoriously challenging predictions of isolated binding sites from low-throughput experiments. Now we can accurately predict *cis*-regulatory modules, conserved clusters of binding sites (**Chapters 13** and **15**), partly based on high-throughput chromatin immunoprecipitation experiments in which tens of millions of DNA segments are sequenced by massively parallel, next-generation sequencers (ChIP-seq, **Chapters 9**, **10**, and **11**). These spectacular developments have allowed for the genome-wide mappings of tens of thousands of transcription factor binding sites in yeast, bacteria, mammals, insects, worms, and plants.

Please also note the no less spectacular failures in many laboratories around the world. Having access to chromatin immunoprecipitation, next-generation sequencing, and software is no guarantee for success. The productive and creative use of computational and experimental tools requires a high-level understanding of the underlying biology, the technological characteristics, and the potential and limitation of statistical and computational solutions. This is the *raison d'être* of this volume, guiding scientists of all disciplines through the jungle of regulatory regions, ChIP-seq, about 200 motif discovery tools and others. As in previous volumes of the series *Methods in Molecular Biology*^{TM}, we help readers to understand the basic principles and give detailed guidance for the computational analyses and biological interpretations of transcription factor binding. We disclose critical practical information and caveats that may be missing from research publications. This volume serves not only computational biologists but experimentalists as well, who may want to understand better how to design and execute experiments and to communicate effectively with computational biologists, computer scientists, and statisticians. **Chapter 1** helps readers to find their way in the maze of resources by a high-level overview of the computational, biological, and some experimental solutions of transcription factor binding. **Chapter 1** highlights other units in this volume and discusses some of the issues not covered.

Why are there so many failed experiments and analyses? Consider, for an example, ChIP-seq, where background noise accounts for more than half of the sequencing reads. Potentially, this may lead to a vast array of false-positive observations. Careful investigators, however, can apply kernel-based density estimates and other background modeling and correction methods to find significantly enriched signals in such noisy observations (**Chapters 9** and **10**). Density estimates are followed by improved peak calling with controlled false discovery rate (**Chapter 10**). Another problem is that ChIP-seq peaks are tens to hundreds of times wider than the footprint of the transcription factor on the DNA. The highest peaks often come from amplification and sequencing bias,

not from a bona fide biological signal (**Chapter 1**). These serious issues mandate the identification of shared, short, and variable DNA motifs, representations of variable binding sites, from moderate-to-low resolution ChIP-seq data using computational motif discovery algorithms. On the other hand, false negatives are also abundant. Consider the temporary nature of regulation, which responds to temporary environmental and internal stimuli. Therefore, a site is typically bound only at a fraction of time, easily missed by snapshot techniques like ChIP (**Chapter 24**). In order to reduce the number of false positives and negatives, motifs are trained by a wide spectrum of statistical learning methods. In spite of the diverse implementation of these tools, most of them stem from expectation maximization and Gibbs sampling (**Chapters 6**, **7**, and **11**) or support vector machines (**Chapter 13**). The trained tools can find binding sites missed by experiments in the predicted promoter regions (**Chapter 5**), all regulatory regions (**Chapter 4**), or in the whole genome.

In itself, de novo computational motif prediction is still not accurate enough (**Chapter 8**). Confidence levels can be increased greatly by integrating binding site locations with in vitro protein–DNA affinities (**Chapter 12**), evolutionary conserved regions (**Chapters 11**, **14**, and **18**), and transposable DNA elements that propagate binding sites through the genome (**Chapter 14**). Time-delayed co-expression as inferred from large compendia of gene expression experiments also indicates binding sites of shared transcription factors. This enormous wealth of information can be retrieved in computationally efficient ways from diverse databases including OregAnno (**Chapter 20**), PlantTFDB (**Chapter 21**), cis-Lexicon (**Chapter 22**), and genome browsers (**Chapters 1**, **10**, and **22**).

The integrated observations and predictions help us to reconstruct complex, hierarchical, and dynamic transcriptional regulatory networks (**Chapters 23** and **24**). This task demands not only new experiments but also the re-annotation of existing experimental data and computational predictions and ongoing, major paradigm changes for all of us.

Istvan Ladunga

http://www.springer.com/978-1-60761-853-9

Computational Biology of Transcription Factor Binding
Ladunga, I. (Ed.)
2010, XI, 454 p. 102 illus., 9 illus. in color., Hardcover
ISBN: 978-1-60761-853-9
A product of Humana Press