BOOK REVIEWS

Evolution, 59(10), 2005, pp. 2281–2283

EVOLUTION FOR BIOINFORMATICIANS AND BIOINFORMATICS FOR EVOLUTIONISTS[1]

Jianzhi Zhang
Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan 48109-1048
E-mail: jianzhi@umich.edu

Received August 3, 2005.

The past decade has seen an explosion of information in many disciplines of biology, particularly in genetics and genomics. For example, as of June 2005, the total number of nucleotides in the DNA sequences available at the GenBank exceeds 89 billion, and this number doubles approximately every 1.5 years. The analysis of this and other large datasets of biological information often involves professional computational scientists and tools, whereas the interpretation of the obtained results usually requires biologists. Bioinformatics was thus born to bridge the gap between computer science and biology. In essence, it uses computational methods to study biological data.

Bioinformatics is such a popular subject that a simple Google search gives about 20 million hits of web pages. In comparison, *systematics* and *phylogenetics* together gets less than 2.5 million hits, and *evolution*, our favorite subject, receives about 87 million hits. It is thus not surprising that almost every major research university in the United States offers graduate and/or undergraduate bioinformatics degrees, although the first papers with the word *bioinformatics* in abstracts or titles are only 12 years old (Bains 1993; Beltrame and Tagliasco 1993; Franklin 1993). Because ‘‘nothing in biology makes sense except in the light of evolution,’’ bioinformaticians need to be enlightened by evolution too. Paul Higgs of McMaster University in Canada and Teresa Attwood of Manchester University in the United Kingdom have now written a book just for this purpose. As stated in the preface, *Bioinformatics and Molecular Evolution* is written for ‘‘people who want to understand bioinformatics methods and who may want to go on to develop methods for themselves.’’ (p. xi) However, the book can also be used by evolutionists to learn bioinformatics, as it will soon be impossible to circumvent bioinformatics in almost any study of evolutionary genetics.

The book has 13 chapters, starting with two beautifully written introductory chapters. Chapter 1 reviews the revolution in biological information and explains why bioinformaticians should know about evolution, especially molecular evolution. Just to give one reason, the single most used tool in bioinformatics is probably the basic local alignment search tool (BLAST), which is employed to identify DNA or protein sequences that are so similar to a query sequence that the similarity is unlikely to be due to chance. This criterion implies that the similarity is most likely due to common ancestry. Hence, BLAST effectively looks for homologous sequences. How can bioinformaticians not understand evolutionary concepts such as homology, common ancestry, convergence, and divergence? After a review of the basics of molecular biology, chapter 2 uses the physicochemical properties of 20 amino acids as an example to introduce one of the most important techniques in bioinformatics, clustering. This carefully written 12-page example is lucid, yet rich in bioinformatic flavors. Several clustering methods are compared in the context of this simple but real example. I am sure it will be liked by students and professors alike.

The rest of the book is roughly separated into two parts, dealing with molecular evolution and bioinformatics, respectively. The five chapters on molecular evolution begin with basic concepts in population genetics (chapter 3), covering topics such as mutation, drift, coalescence, genealogical trees, selection, adaptive sweeps, and the neutral theory. Although the descriptions are brief, they are accurate and easy to understand. The next chapter (4) discusses models of DNA and protein sequence evolution. Because these models are used frequently in phylogenetic analysis and in database searches, the authors explain in great detail how the models are derived and how the parameters are estimated from real data. The latter subject is often omitted (Li 1997) or only briefly mentioned (Nei and Kumar 2000) in molecular evolution books, probably because it is too bioinformatic. I am glad that this book treats it thoroughly. The book then introduces molecular phylogenetics (chapter 8), covering the commonly used distance, parsimony, likelihood, and Bayesian methods, as well as the bootstrap technique that is used to evaluate the precision (or reliability) of reconstructed phylogenies. The authors apparently favor the Bayesian method with the Markov chain Monte Carlo (MCMC) tree-search algorithm, as they believe that ‘‘it is the best way of making maximum use of the information in the data.’’ (p. 184) However, not every phylogeneticist would agree (Felsenstein 2003). In fact, the excessively high clade-supporting posterior probabilities provided by MCMC are worrisome because they appear to overestimate the probability that the reconstructed clades are true (Douady et al. 2003; Suzuki et al. 2002). Chapter 11 provides additional topics in molecular evolution, including RNA structure and evolution, fitting evolutionary models to sequence data, and several case studies in molecular systematics. Chapter 12 is on genome evolution, discussing genome size, gene content, gene rearrangement, gene gains and losses, and the evolution of organellar genomes. Both authors of the book were trained as biophysicists. As a molecular evolutionist, I read the above chapters with admiration and amazement

[1] *Bioinformatics and Molecular Evolution.* Paul G. Higgs and Teresa K. Attwood. 2005. Blackwell Publishing, Oxford, U.K. xiii + 365 pp. PB $74.95, ISBN 1-4051-0683-2.

because I can find virtually no mistakes. Instead, their explanations of population genetic and molecular evolutionary concepts and methods are succinct and accurate. The carefully picked examples are all real cases and illustrate the concepts and methods quite well.

The remaining six chapters cover bioinformatics methods, starting with an introduction of biological databases (chapter 5). Although sequence databases and sequence-based databases (e.g., protein family and domain databases) are described in detail, other types of biological databases (except protein structure databases) are not well introduced. For instance, there is only one sentence about the many protein-interaction and metabolic-pathway databases. The next chapter discusses algorithms of sequence alignment (chapter 6). Sequence alignment is important in many phylogenetic and evolutionary studies and is one of the first steps in sequence analysis. But most molecular evolution texts (e.g., Li 1997; Page and Holmes 1998; Nei and Kumar 2000) treat it rather casually and briefly. This book provides detailed explanations of alignment algorithms that are accessible to molecular evolutionists. Chapter 7 is on searching sequence databases, which molecular evolutionists do daily. The introduction of the algorithms of several popular searching programs such as BLAST and FASTA is very helpful. I bet many molecular evolutionists do not know the exact meanings of *score* and *E value* in the BLAST output screen. The book provides comprehensive explanations and practical recommendations for the use of these statistics. Chapter 9 discusses pattern recognition in protein families, which is used to identify conserved sequence motifs from a larger number of sequences. These motifs can then be used to predict protein functions and to classify proteins. Although pattern recognition is an important task for bioinformaticians, it is rarely used by molecular evolutionists. But the example of classifying the large number of G-protein coupled receptors from eukaryotes makes it clear that we molecular evolutionists can benefit from the use of pattern recognition tools in analyzing distantly related proteins. Chapter 10 introduces probabilistic models and machine learning methods for pattern recognition. Although Bayesian statistics and likelihood ratio tests are known to many molecular evolutionists, other topics such as neural network models and hidden Markov models are relatively new. But the introductions are accessible and I believe molecular evolutionists will appreciate these methods. Chapter 13 is about many ’omes other than the genome. For example, *transcriptome* refers to the complete set of mRNAs in a cell, *proteome* refers to the complete set of proteins in a cell, and *metabolome* refers to the complete set of chemicals involved in metabolic reactions in a cell. The chapter introduces high-throughput molecular techniques for obtaining ’ome-wide datasets and presents basic statistical methods used to process these data.

I have two disappointments after reading the book. First, the molecular evolution topics covered in the book are too limited. The authors discuss mostly molecular phylogenetics and the techniques and concepts indispensable for tree-making (e.g., sequence alignment and substitution models), but there are many other interesting subjects that are both important and stimulating for bioinformaticians. For example, evolution by gene duplication is one of the general principles of evolution (Ohno 1970; Zhang 2003), but the book has only one paragraph on it, without even providing an example or a reference. In fact, gene duplication and gene family evolution has been an energetic field since a decade ago because of the availability of genomic data and bioinformatic analysis of these data. One testimony is the dramatic increase of the citation of Susumo Ohno’s (1970) classic book *Evolution by Gene Duplication* since the mid-1990s. Another example of interesting topics left largely untouched is the relative roles of natural selection and genetic drift in evolution (Kimura 1983). Bioinformatic analysis of genomic data has the potential to answer this important question and therefore it should be introduced to bioinformaticians. Unfortunately, although the authors describe some population genetic concepts in chapter 3 as prerequisites for evolution, they almost never use them in later chapters of the book. Processes and mechanisms of evolution are not emphasized in this book, making the evolutionary topics less interesting and attractive than they could be and should be.

My second disappointment is that the book emphasizes almost exclusively the analysis of sequence data, either of proteins or DNA. Although it is true that sequence-based analysis is the best developed, recent years have seen many interesting evolutionary and bioinformatic analyses of other types of data, such as protein interaction data, transcriptional regulation data, gene expression data, and single-gene-deletion fitness data. Although some of these data are mentioned in the book, the treatment is casual and descriptive. I believe that these topics can be at least as interesting as, or even more interesting than, sequence-based analysis, if they are presented in an evolutionary context and in a mechanistic perspective. The book could also have introduced the network thinking and systems biology thinking, which are starting in the molecular biology and evolution communities (Kitano 2002; Proulx et al. 2005). Although the correctness and usefulness of these views are yet to be proven, the fresh perspectives will allow both evolutionists and bioinformaticians to think out of box and look for new horizons through cross-disciplinary brainstorming in this postgenome era. These limitations notwithstanding, the authors should be applauded for making this book highly accessible to both evolutionists and bioinformaticians.

Literature Cited

Bains, W. 1993. Bioinformatics in Europe: the federation strikes back. Trends Biotechnol. 11:217–218.

Beltrame, F., and V. Tagliasco. 1993. Confocal microscopy and cellular bioinformatics. Cytotechnology 11(Suppl 1):S72–S74.

Douady, C. J., F. Delsuc, Y. Boucher, W. F. Doolittle, and E. J. Douzery. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol. Biol. Evol. 20:248–254.

Felsenstein, J. 2003. Inferring phylogenies. Sinauer, Sunderland, MA.

Franklin, J. 1993. Bioinformatics changing the face of information. Ann. NY Acad. Sci. 700:145–152.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge Univ. Press, Cambridge, U.K.

Kitano, H. 2002. Computational systems biology. Nature 420: 206–210.

Li, W. H. 1997. Molecular evolution. Sinauer, Sunderland, MA.

Nei, M., and S. Kumar. 2000. Molecular evolution and phylogenetics. Oxford Univ. Press, New York.

Ohno, S. 1970. Evolution by gene duplication. Springer-Verlag, Berlin.

Page, R. D. M., and E. C. Holmes. 1998. Molecular evolution; a phylogenetic perspective. Blackwell Science, Oxford, U.K.

Proulx, S., D. Promislow, and P. Phillips. 2005. Network thinking: next term in ecology and evolution. Trends Ecol. Evol. 20: 345–353.

Suzuki, Y., G. V. Glazko, and M. Nei. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc. Natl. Acad. Sci. USA 99:16138–16143.

Zhang, J. 2003. Evolution by gene duplication: an update. Trends Ecol. Evol. 18:292–298.

Book Review Editor: D. Futuyma