

PREFACE

Biological and biomedical sciences are becoming more interdisciplinary, and scientists of the future need interdisciplinary training instead of the conventional disciplinary training. Just as Sean Eddy (2005) wisely pointed out that sending monolingual diplomats to the United Nations may not enhance international collaborations, combining strictly disciplinary scientists trained in either mathematics, computational science or molecular biology will not create a productive interdisciplinary team ready to solve interdisciplinary problems.

Molecular biology is an interdisciplinary science back in its heyday, and founders of molecular biology were often interdisciplinary scientists. Indeed, Francis Crick considered himself as “a mixture of crystallographer, biophysicist, biochemist, and geneticist” (Crick, 1965). Because it was too cumbersome to explain to people that he was such a mixture, the term “molecular biologist” came handy. To get the crystallographer, biophysicist, biochemist, and geneticist within himself to collaborate with each other probably worked better than a team with a crystallographer, a biophysicist, a biochemist and a geneticist who may not even be interested in each other’s problems.

Bioinformatics was born in response to the interdisciplinary demand of modern biological and biomedical sciences as a joint effort by among mathematicians, computational scientists and biologists of all colors and shades. It is a peculiar branch of science. A conventional branch of science such as quantum mechanics in physics or population genetics in biology will typically have a few classic publications laying down its theoretical foundation, delineating its boundary and interface with other related sciences, formulating its central questions, and highlighting the spectacular views within and beyond that particular mansion of science. Once the mansion has been skeletally constructed, subsequent works will only serve to

beautify the mansion but will not alter the general structure of the mansion which remains easily recognizable by people within the mansion and those in its neighbourhood. There is little controversy as to how the mansion looks, even when viewed from different perspectives.

Bioinformatics is different. I have been told that bioinformatics was not built by one or a few visionary giants of science, and that it is not even a single mansion. Instead, it is the Wild West before the law arrives (Eddy, 2005), dotted by a large number of trailer houses or even temporary tents that have been built and found workable elsewhere in the kingdom of science. Many people living in this rough terrain do not know where they belong but, after living here for some time, found it necessary to give the dwelling a label. When someone murmured the word “bioinformatics”, everyone thought it a godsend and the town of bioinformatics was born, and the inhabitants begin to call themselves bioinformaticians.

Bioinformaticians differ dramatically in their views and their descriptions of their town. This is partially reflected in the flagship journal of the field, *Bioinformatics*. Most papers in the journal were treated by conventional biologists as Wild West stories and ignored with a passion, except for a few that get a great deal of attention and citation by proclaiming the finding of gold. The only consensus among bioinformaticians seems to be that bioinformatics deals with very big computational problems. However, when asked about what the very big problems are, most bioinformaticians, to paraphrase Peter Medawar, will become instantly solemn and shifty-eyed, solemn because they think that they have something profound to declare, and shifty-eyed because they really have nothing to declare.

Such a perception of bioinformatics is witty but unfair, because bioinformatics does have a root to trace to and a central theme with a focus. For many years, a challenging question near and dear to the mind of many leading biologists is how living cells work. A living cell is a system with cellular components interacting with each other and with extracellular environment, and these interactions determine the fate of the cell, e.g., whether a stem cell is going to become a liver cell, a brain cell, or a cancerous cell. It then became quite obvious that, to understand how living cells work, those cellular components and their interactions would need to be identified and characterized. The most important cellular components happened to be universally acknowledged to be the genome, the transcripts and the proteins. The characterization and analysis of these three types of cellular components leads to genomics, transcriptomics and proteomics that jointly drive the development of bioinformatics.

Genomics leads to two developments. The first is to allow a much faster identification of proteins by combining mass spectrometry data with genomic databases. Second, genomic sequences have enabled SAGE

(sequential analysis of gene expression) and microarray technology which have spawned transcriptomics which is a synonym of functional genomics. Biologists can now routinely monitor the gene expression at the genomic scale over time or compare gene expression between control cells and treatment cells or along the developmental path of a particular cell type.

There are two major problems with the transcriptomic data. The first is that the relative abundance of transcripts as characterized by SAGE or microarray experiments is not always a good predictor of the relative abundance of proteins, yet proteins are true workhorses in the cell. Many proteins that are produced as a result of alternative splicing and posttranslational modification will not reveal their mystery in our analysis of transcriptomic data. It should be quite clear that, in order to characterize the cellular components and their interactions, one needs the corroboration of proteomic, genomic and transcriptomic data.

At this point, one is tempted to conclude that bioinformatics has three facets labelled proteomics, genomics and transcriptomics and that it deals mainly with characterizing cellular components and their interactions. But bioinformatics goes beyond this. Genes and genomes have evolved from time immemorial, as do interactions among genes and gene products. The genomic change is particularly well exemplified by the infectious diseases caused by the influenza viruses, the SARS virus, and HIV, all evolving quickly as a result of mutation, recombination and selection. Studying the dynamic nature of genes and genomes, tracing their phylogenetic relationships and reconstructing their ancestral states allow biologists to gain the advantage perceived by Aristotle thousands of years ago, i.e., “He who sees things grow from the very beginning has the most advantageous view of them.” For this reason, molecular evolution is now an essential component of bioinformatics.

Many books have now been written on bioinformatics. They tend to fall on two extremes. In one extreme are books featuring computational details with a great deal of mathematics (e.g., Pevzner, 2000), while in the other extreme one finds books treating bioinformatics mostly as a giant black box (e.g., Baxevanis and Ouellette, 2005). The former is for computational scientists and mathematicians who, after reading the book, will remain computational scientists and mathematicians. The latter is for biologists who, after reading the book, will remain biologists. Such books often have limited contribution to creating interdisciplinary scientists needed in modern biological and biomedical sciences.

Most biologists cannot appreciate the beauty of mathematics without having the equations rendered to numbers. Remarkably, neither can mathematicians and computational scientists appreciate the beauty of biology without having the cells and bugs rendered to numbers. This book is

my effort to render both mathematical equations and biology to numbers. It is aimed at creating truly interdisciplinary scientists to prosper in the Wild West of bioinformatics.

Although the book covers bioinformatics methods at a level more advanced than most other bioinformatics books, the extensive numerical illustration of these methods should make it accessible to most senior undergraduate students and graduate students majoring in science and software engineering. An additional advantage of using it as a textbook is that nearly all algorithms in the book are implemented in a free and user-friendly computer program (DAMBE).

Practising biologists with reasonably good programming skills should be able to implement most algorithms themselves and check the output against numerically illustrated examples in the book. They should soon find such learning experience intellectually rewarding and mentally satisfying. Some of them might even be pleasantly surprised to learn that they can quickly create a much needed computational method that their computer technicians have failed to create in a whole year.

I have tried my best to “make everything as simple as possible, but not simpler”. While most numerical illustrations of advanced computational algorithms in this book are toy examples, they require only simple extensions to tackle real data. To paraphrase the late C. C. Li, it is not necessary to create a rainbow spanning the sky to demonstrate how a rainbow forms – a small one is convincing enough.

“Please read the book”.

Bioinformatics and the Cell
Modern Computational Approaches in Genomics,
Proteomics and Transcriptomics

Xia, X.

2007, XVI, 350 p., Hardcover

ISBN: 978-0-387-71336-6