1 Introduction

Topics like biotechnology, genetics, and bioinformatics frequently find their way into today's headlines. This trend, often seen as the beginning of a new era, was initiated by a seminal discovery more than fifty years ago, the discovery of the DNA double helix structure by Watson and Crick in 1953. The development in molecular biology continued to grow steadily from then, and it became more and more important to the public, for example, with the launch of the Human Genome Project in the 1990s and with the presentation of the cloned sheep "Dolly." In 2000, scientists eventually announced the complete sequencing of the human genome.

Let us take a closer look at the task of DNA sequencing. The goal here is to determine the sequence of nucleotides, which are the elementary building blocks in the human DNA, i.e., in the molecule storing our hereditary information. From the viewpoint of informatics, we are looking for a string made up from letters representing these nucleotides. Methods for reading their sequence have already been known since the 1980s, but the length of DNA molecules that can be sequenced using these methods is severely restricted. Current methods enable us to read about 1 000 consecutive nucleotides. Nevertheless, we are mainly interested in DNA molecules consisting of hundreds of thousands of them. How can we reach our goal of sequencing in these cases? One possible approach is the following: We generate a multitude of copies of the DNA molecule of our interest. We randomly break each of these copies into fragments. With high probability, the resulting fragments from different copies overlap with each other. Ideally, these fragments are sufficiently short for direct sequencing. Having performed these sequencing operations, we are left with many string fragments of which we know that they occur as substrings in the DNA sequence, and that these fragments may overlap with each other. But we have no clue how to combine these fragments to achieve the complete DNA sequence, since their order was lost in this process. Typically, we have to reorder thousands of string fragments, a task we are not capable of doing by hand. This is the point where informatics comes in. First of all, it gives us the appropriate tools for managing the data, i.e., the set of string fragments.

Moreover, it helps us to formalize how a clever reconstruction of the DNA sequence would look. One possible example of such a formalization could be to search for the shortest string containing all fragments as substrings. Then we could try to solve the formalized problem with computational methods. But this task is not as easy as it may seem. Due to the enormous number of possible fragment orderings, we cannot rely on computational power alone; even this very intuitive and apparently simple problem leaves us with a great algorithmic challenge. Furthermore, even if we somehow succeed in finding this shortest string containing all fragments, the question remains open as to whether this string really coincides with the prospected DNA sequence, or whether we have to look for some refined model incorporating further aspects of the problem.

In this book, we will deal with questions of this type. In particular, we will describe how to derive a formal model from a biological problem, and how to find a solution to this formal model algorithmically.

Research problems in molecular biology typically do not arise with a given completely formal description. In contrast, one has on the one hand a rather concrete idea of the desired result, but on the other hand only a vague intuition of how this goal could be achieved. Different biological or biotechnological approaches or methods in use can lead to the need for completely different computational methods. Understanding the biological methods and approaches in detail is in many cases possible for biologists only, but knowledge about the basic principles and connections is very helpful for the computer scientists dealing with these kinds of problems. Such knowledge alone enables them to understand the real problems at hand and to possibly suggest modifications to the methods that could help to make the underlying computational tasks easier.

One major concern is that all biological data is inherently inexact. Experimental methods are always error prone, and these errors have to be taken into account in further processing steps. On the other hand, every *solution* derived by computational methods for some biological problem is only a *solution hypothesis* in reality. Only by further investigations does it become clear whether such a hypothesis really induces a biologically relevant solution or whether different aspects of the original problem have to be incorporated into the formal model to gain modified or enhanced solution hypotheses.

To be treated using algorithmic methods, a biological problem has to be transformed into a formal model, i.e., into a formal specification of the problem identifying in particular the data at hand and the desired result. Without such a formal model, it remains unclear how to reach the desired goal using computational methods. In this context, each developed model has to be evaluated according to its ability to describe the relevant real-life aspects of the given biological problem. For example, a common model of the DNA molecule consists of a description of the linear sequence of the nucleotides in terms of a string, as described above. In most cases, this model is sensible and offers a convenient possibility for further processing using computational methods.

But a DNA molecule is no string! For instance, if we want to examine the spatial properties of a DNA molecule, modelling it as a string is clearly not sufficient, and we have to look for new or enhanced models.

When we eventually succeed in finding an appropriate formal model for our biological research problem, we can start to look for algorithmic approaches to solving this formal problem. Even then, we should take the biological realities as a guideline for our examinations. As an example, we will again consider the string model of DNA molecules. Although strings can in general be composed of arbitrarily many different letters, we do not have to take this into account for our considerations, since the DNA is made up of only *four* different nucleotides.

With this reasoning, we will try throughout this book to proceed using the scheme

problem – model – algorithm

This means that we will, for every biological problem, describe the corresponding formal model, or even several models, and discuss their advantages and shortcomings. Only after that will we examine the algorithmic properties of the resulting formal models. Nevertheless, the focus of our attention will lie on the formal models and the algorithmic approaches for finding solutions for them, but not without keeping an eye on the underlying biological applications.

The complete spectrum of computational methods is used for investigating problems from molecular biology. In particular, methods from the fields of database management, statistics, and algorithmics are used. In this book, we will focus on the algorithmic aspects. In the subsequent chapters, we will also present some of the basic concepts from algorithm theory, for example, algorithm design methods like dynamic programming, divide and conquer, backtracking, branch and bound, and many more. We will in particular consider the concept of approximation algorithms as it arises in the field of combinatorial optimization. But all the presented methods will be embedded into the context of actual biological problems and are illustrated using concrete examples.

This book is designed as a self-contained textbook. Its goal is to explain the basic principles of algorithmic bioinformatics to students, and to help lecturers prepare introductory courses on the subject. This book is primarily targeted at graduate and advanced undergraduate students of computer science and at life sciences students interested in algorithmics. To successfully read this book, only very basic knowledge of data structures and algorithms is assumed. We try to cover a multitude of problems, and to present an overview of the models and the methods used for solving them. In this way, we hope to impart a solid basis of knowledge enabling readers to build on and intensify their studies in their respective research fields. Nevertheless, we have to mention that it is impossible for an introductory textbook to cover all topics in depth; but we hope that our choice of topics stimulates the reader's interest.

This book is divided into three parts. The first part serves as an introduction to the field of bioinformatics. After a short overview of the biological background and the basic notions of algorithmics, we will deal here with algorithmic methods concerning strings, for instance, string matching algorithms and methods for computing the similarity of strings. The second part is devoted to the field of DNA sequencing, which provides an important source of data for further investigations. In this part, we deal with generating physical maps and with different approaches and models of the actual DNA sequencing process. This is followed by the third part, in which we analyze a multitude of various problems arising in molecular biology. Among other problems, we deal with signal finding in DNA sequences, phylogenies, haplotyping, and the computation of spatial molecular structures. Furthermore, each chapter contains, besides those sections in which we present the problems and their solutions in detail, one section summarizing the presented material. This overview of results assists the readers in self-checking their understanding of the presented material. Moreover, every chapter concludes with a section pointing to the papers and books we used for preparing the chapter as well as to additional literature, and thus directs the interested reader to sources for further study.

We have tried to motivate and to describe the presented topics as accurately as possible. We appreciate any comments and suggestions as well as any information about remaining errors. Please note the following website:

`http://www.ite.ethz.ch/publications/bioinformatics`

Finally, we hope that we will be able to create interest and excitement for the field of bioinformatics. Please enjoy reading!