Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

European Journal of Operational Research 161 (2005) 585–597

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

www.elsevier.com/locate/dsw

Invited Review

Selected combinatorial problems of computational biology

Jacek Błażewicz *[,1], Piotr Formanowicz, Marta Kasprzak

Institute of Computing Science, Poznań University of Technology, ul. Piotrowo 3A, 60-965 Poznań, Poland
Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznań, Poland

Received 9 April 2003; accepted 3 October 2003
Available online 14 May 2004

Abstract

Recently we observe a great breakthrough in biology connected with the studies on genomes. These achievements would be impossible without an input from other sciences, combinatorial optimization being one of them. This study is devoted to a presentation of the most important (to our opinion) area of the computational biology, mostly connected with DNA studies, where combinatorial optimization impact was clearly visible. They include: sequencing DNA chains, assembling them, genome mapping and sequence comparison.

Keywords: DNA sequencing by hybridization; Sequence assembling; Genome mapping; Sequence alignment

1. Introduction

In recent years several spectacular events connected with genome studies have occurred, reading the human genome being one of them. Biological sciences got a great impact on many aspects of the everyday life. Since a discovery by Watson and Crick their double helix model of a DNA chain [40], biology has made a great progress in understanding foundations of life. The progress would have not been possible, however, without a help from other areas of science. Here, let us mention physics, chemistry and last but not least computing sciences. We will concentrate on the impact on biology of the latter science, especially its part connected with *operational research* and *combinatorial optimization*. Since ties between biology and combinatorial optimization may be found in very many fields of biology, we will concentrate on the most important (to our opinion) examples of the application of combinatorial optimization methodology in modeling and solving problems arising in the context of the computational biology. They include: sequencing DNA chains, assembling them, genome mapping and sequence comparison. The aim of the paper was to present these problems in a way that is interesting to those working in the field, as well as non-specialists which may

[*] Corresponding author. Address: Institute of Computing Science, Poznań University of Technology, ul. Piotrowo 3A, 60-965 Poznań, Poland.
E-mail address: blazewic@put.poznan.pl (J. Błażewicz).

[1] The paper was written while the first author was at Laboratoire Leibniz, Universite Joseph Fourier, whose help is greately appreciated.

0377-2217/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2003.10.054

become interested by the new exciting applications of operational research. Thus, an up to date review of research in the field is complemented by many examples allowing one for a better understanding of the concepts introduced. An interested reader is referred also to the recent monographs, which describe the above and other problems in a greater detail: [18,20,30,33,39].

The organization of the paper is as follows. Section 2 contains a biological primer. Section 3 discusses the DNA sequencing problem, while Section 4 presents the assembling one. Genome mapping and sequence comparison are considered, respectively, in Sections 5 and 6.

2. Biological primer

Biology is a science that provides an understanding of the nature of all living things at different levels – from molecules to cells, individuals and populations. It is accepted that all living organisms are composed either of a single cell or a collection of them. At this stage of development a primary interest of biology lies in molecules and cells, thus, we have molecular biology. With the aid of computing science models and tools (combinatorial optimization being one of its main components) it creates so called computational molecular biology.

One of the main objects of the study of computational biology are *DNA chains*, coding genetic information of living organisms. DNA is a *string* composed of letters (*nucleotides*) being members of the alphabet {A, C, G, T}. Short single-stranded DNA molecules are called *oligonucleotides*. The entire DNA of the organism is called its *genome* and its length may reach billions of nucleotides (or base pairs). The same genome is contained in each cell of a given living organism (e.g. human beings have trillions of cells). DNA appears in a form of a *double helix*, i.e. a double strand, where A in one chain can be bound to T only, and C to G, respectively. This fundamental law of a DNA construction was discovered by Watson and Crick [40]. Knowing, thus, one strand of a DNA helix, the second (*complementary*) can be easily reconstructed. What is more, this property of a single strand (trying to bind to a complementary strand) called a *hybridization*, may be used in laboratories in many processes leading e.g. to a reconstruction of an unknown chain.

The genetic information contained in DNA is then used, as stated by the Central Dogma of Molecular Biology, to produce *RNA* and ultimately *proteins*. The latter are the main construction material of living organisms, deciding of their functioning as well.

In the following, we will concentrate on DNA analysis and reconstruction, since these processes require a considerable input from the combinatorial optimization side. We will consider combinatorial problems connected with DNA reading, sequence comparison and phylogenetic analysis.

Reading sequences of genomic DNA is usually a starting point for further molecular biology research. But reading DNA sequences itself is not a trivial task and may involve some sophisticated procedures. It follows from the fact that it is not possible to read a sequence of nucleotides directly, e.g. using a microscope. Hence, some indirect methods have to be used. The process of reading the sequence of a genome is usually divided into three stages: *mapping*, *assembling* and *sequencing*.

The process of reading a long piece of DNA usually starts with cutting it into smaller pieces of size about 100 000–1 000 000 nucleotides. (Let us note that biological experiments take effects usually on millions of copies of given types of molecules.) During the cutting process the information about the order of these pieces is lost. But, as one can guess, it is necessary to recover this information and this is done by mapping procedures. When mapped, the fragment is picked up and cut into much smaller parts of length about 40 000 nucleotides. Again, mapping is necessary to recover the order of the pieces obtained by the second-level cutting [20].

Since sequencing methods allow determining sequences of lengths not greater than 1000 nucleotides, fragments of lengths about 40 000 base pairs cannot be directly sequenced. So, it is necessary to break them into pieces of appropriate lengths. This is done randomly. At this stage one gets DNA fragments suitable for *sequencing*. After sequencing the short fragments are tried to be