Available online at www.sciencedirect.com
ScienceDirect
Journal of Visual Languages and Computing
19 (2008) 291–301

Journal of
Visual Languages
& Computing

www.elsevier.com/locate/jvlc

Visual software tools for bioinformatics

Timothy Arndt*

Department of Computer and Information Sciences, Cleveland State University, 2121 Euclid Avenue, Cleveland, OH 44115-2124, USA

Received 11 June 2007; accepted 15 June 2007

Abstract

Bioinformatics is the application of techniques from computer science, statistics and mathematics to problems in molecular biology. This interdisciplinary approach is rapidly revolutionizing biology. A survey of software tools for bioinformatics is presented. A special emphasis is placed on the visual aspects of these tools. The most important visualization tasks in bioinformatics are data sequence visualization and visualizing protein structures. The visualization of interactions between molecules in a metabolic pathway or network is an emerging area. Many important visualization techniques have yet to be applied in this application area.

Keywords: Bioinformatics; Software tools; Reviews

1. Introduction

Bioinformatics has been defined as the application of information technology (computer science, mathematics and statistics) to the management of biological information. In particular, bioinformatics has been widely associated with molecular biology that is largely concerned with the study of three types of molecules—DNA, RNA and protein. The *central dogma* of molecular biology describes how the information stored in DNA is transcribed into RNA and then translated into protein. Each of these three molecules is a polymer—a string of simpler units, nucleotides in the case of DNA and RNA, amino acids in the case of protein. Each nucleotide contains one of four bases—adenine (abbreviated A),

*Tel.: +1 216 687 4779; fax: +1 216 687 5448.
E-mail address: arndt@grail.cba.csuohio.edu

1045-926X/$ - see front matter
doi:10.1016/j.jvlc.2007.06.001

cytosine (C), guanine (G) and thymine (T). Uracil (U) takes the place of thymine in RNA. There are 20 naturally occurring amino acids. Each amino acid can be specified by either a three letter or a one letter code. For example, tryptophan is specified by either Trp or W. It is easily seen that the one letter code is more appropriate for computer processing.

The DNA molecule has the famous double helix structure in which each base from one strand of the double helix pairs with a base from the other strand. An A base pairs only with a T base, while a C base pairs only with a G base. Due to this, given the sequence of one of the strands, we can infer the sequence data for the complementary strand. Thus, a DNA molecule can be specified by giving the sequence of one the strands, for example, AAACGTC etc. The story is a bit more complex for RNA and proteins, which are single stranded. We still can usefully characterize the molecule by giving the sequence data (a string of bases for RNA, a string of amino acids for protein), however this does not completely characterize the molecules, since they can fold into irregular shapes which are functionally important. For these molecules, the sequence data is referred to as the primary structure while the secondary structure is the three-dimensional form of local segments of the polymer. Typical local structures for proteins are alpha helices and beta sheets while the stem-loop is a typical RNA secondary structure. The tertiary structure of a protein is its three-dimensional structure given by the atomic coordinates, while quaternary structure is the arrangement of multiple folded proteins in a protein complex. Fig. 1 below shows the secondary structure of the myoglobin protein which contains several alpha helices and random coils, but no beta sheets. The visual representations for the two secondary structures is typical.

One of the most important tasks of a bioinformatics tool is to perform sequence alignment. Given two different but related sequences (of possibly different lengths), the tool attempts to find the best match between them. In Fig. 2, the alignment between two

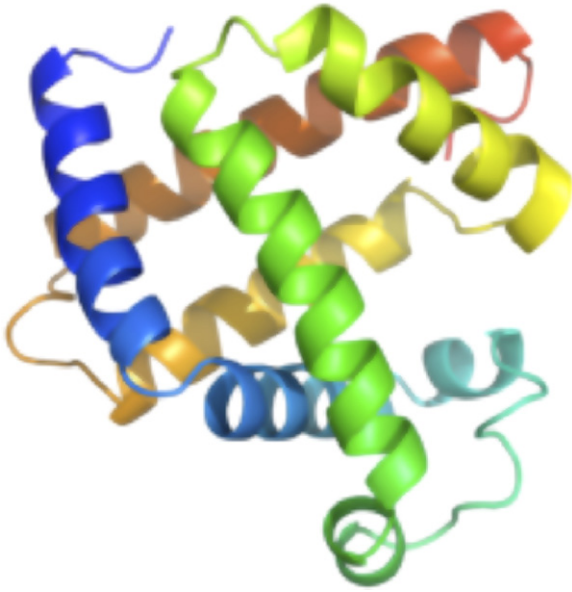

Fig. 1. The secondary structure of the myoglobin protein.