Web Tools for Molecular Biological Data Analysis

Denise Fukumi Tsunoda
Centro Federal de Educação Tecnológica do Paraná, CEFET/PR, Brazil

Heitor Silvério Lopes
Centro Federal de Educação Tecnológica do Paraná, CEFET/PR, Brazil

Ana Tereza Vasconcelos
Laboratório Nacional de Computação Científica, LNCC, Brazil

INTRODUCTION

Bioinformatics means solving problems arising from biology using methods from computer science. The National Center for Biotechnology Information (*www.ncbi.nih.gov*) defines bioinformatics as:

"…the field of science in which biology, computer science, and information technology merge into a single discipline...There are three important sub-disciplines within bioinformatics: the development of new algorithms and statistics with which to access relationships among members of large data sets; the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures; and the development and implementation of tools that enable efficient access and management of different types of information."

There are many sub-areas in bioinformatics: data comparison, data analysis, DNA assembly, protein structure prediction, data visualization, protein alignment, phylogenetic analysis, drug design, and others.

The biological data (sequences and structures) are naturally very large. In addition, the number of records in the biological databases is increasing every year because of intensive research in the field of molecular biology. Analysis of this overwhelming amount of data requires intelligent bioinformatics tools in order to manage these data efficiently.

During the past two decades, the world has witnessed a technological evolution that has provided an unprecedented new medium of communications to mankind. By means of the World Wide Web, information in all forms has been disseminated throughout the world. Since the beginning, research in bioinformatics primarily used the Internet due to the fast information dissemination it allows, at essentially no cost.

This article aims to discuss some bioinformatics Web tools, but given the accelerated growth of the Web and the instability of the URLs (Uniform Resource Locators), an Internet search engine should be used to identify the current URL.

BACKGROUND

Living organisms possess entities called *genes* that are the basic inherited units of biological function and structure. An organism inherits its genes from its parents, and relays its own genes to its offspring.

Molecular biologists, in the later half of the 20th century, determined that the gene is made of DNA (deoxyribonucleic acid)—that is, DNA is the heredity material of all species. More than 50 years ago, Crick and Watson (1953) discovered the *double helix* structure of DNA and concluded that this specific form is fundamental to DNA's function.

Each strand of the DNA double helix is a polymer (a compound made up of small simple molecules) consisting of four elements called *nucleotides*: *A*, *T*, *C*, and *G* (for adenine, thymine, cytosine, and guanine). The two strands of DNA are complementary: when a *T* resides on one strand, an *A* occupies the corresponding position on the other strand; when there is a *G* on one strand, a *C* occupies the corresponding position on the other. The sequence of nucleotides encodes the "instructions" for forming all other cellular components and provides a template for the production of an identical second strand in a process called replication.

From a computer scientist's point of view, the DNA is information storage and a transmission system. Like the binary alphabet {0,1} used in computers, the alphabet of DNA {*A, T, C, G*} can encode messages of arbitrary complexity when encoded into long sequences.

The decoding of the genetic information is carried out through intermediary RNA (ribonucleic acid) molecules that are transcribed from specific regions of the DNA. RNA molecules use the base uracile (U) instead of a thymine. RNA is then translated into a protein—a chain assembled from the 20 different simple amino acids. Each consecutive triplet of DNA elements specifies one amino acid in a protein chain. Once synthesized, the protein chain folds—according to the laws of chemistry/physics—into a specific shape, based on the properties and order of its amino acids. The structures of a protein can be viewed hierarchically (Lehninger, Nelson & Cox, 2000): primary (linear amino acid sequence), secondary (local sequence elements with well-determined regular shape like á-helices and â-strands), tertiary (formed by packing secondary structures into one or several compact globular units), and quaternary (combination of tertiary structures).

BIOINFORMATICS WEB TOOLS

Sequence Analysis

There is a known relationship between sequence and structure of proteins, since proteins with similar sequences tend to have similar three-dimensional structures and functions. Sequence alignment methods are useful when it is necessary to predict the structure (or function) of a new protein whose sequence has just been determined. Therefore, alignment provides a powerful tool to compare two (or more) sequences and could reflect a common evolutionary origin.

ClustalW is a general purpose multiple sequence alignment program for DNA or proteins (Higgins et al., 1994). ClustalW currently supports seven multiple sequence formats that are detailed (including examples) in the *ClustalW Services Help Menu*: NBRF/PIR, EMBL/SwissProt, FASTA, GDE, ALN/ClustalW, GCG/MSF, and RSF. It produces biologically meaningful multiple sequence alignments of divergent sequences and calculates the best match for the selected sequences, considering individual weights, amino acid substitution matrices—like PAM (Altschul, Gish, Miller, Myers & Lipman, 1991) or Blosum (Henikoff & Henikoff, 1992)—and gap penalties (Apostolico & Giancarlo, 1998). After the identities, similarities and differences can be seen. ClustalW is freely available on the Internet, either as a Web-based tool or for downloading.

Another tool, T-Coffee (Notredame, Higgins & Heringa, 2000) is more accurate than ClustalW for sequences with less than 30% identity, but much slower. The T-Coffee input must have from 3 to 30 sequences (or 10,000 characters) in the FASTA format. The submission form is simple, but it does not allow user-selected options.

Cinema—Colour INteractive Editor for Multiple Alignments—is a program for sequence alignment that allows visualization and manipulation of both protein and DNA sequences (Parry-Smith, Payne, Michie & Attwood, 1997). It is a complete package in Java, locally installed, that runs on most platforms. This tool allows upload of an alignment from a local computer to the Cinema server. The input file must be in a PIR format and may then be imported into Cinema via the *Load Alignment File* option.

Structural Analysis

The Dali—Distance mAtrix aLIgnment—server (Holm & Sander, 1994) is a network service for comparing three-dimensional (3D) protein structures. Once the coordinates of a query protein structure is submitted, Dali compares them against those in the Protein Data Bank (PDB). The input file must be in the PDB format (Berman et al., 2000) and can be submitted by e-mail or interactively from the Web. The input options are *disabled*. The results are mailed back to the user. In favorable cases, comparing 3D structures may reveal biologically interesting similarities that are not detectable in primary sequences. There is a Dali database built based on exhaustive all-against-all 3D structure comparison of protein structures currently in the Protein Data Bank (PDB). The classification and alignments are continuously updated using the Dali search engine.

The Macromolecular Structure Database tool (MSD) (Golovin et al., 2004) allows one to search the active site database based on ligand or active site information. The PDB contains a significant number of protein structures that have ligands bound which are often more highly conserved across a functional family than the overall structure and fold of the macromolecule. The target of the search can be based on an uploaded file. It is possible to limit the scope of a search using restrictions based on author, keywords, experiment, resolution, and release date. Results of the search are presented in a list of PDB ID codes that can be analyzed further or viewed within a structure viewer like Rasmol (Sayle & Milner-White, 1995)—a program that intends the visualization of proteins, nucleic acids, and small molecules.

Swiss-Model (Schwede, Kopp, Guex & Peitsch, 2003) is a server for automated comparative modeling of 3D protein structures, and provides several levels of user interaction using a Web-based interface: in the first approach mode, only an amino acid sequence is submitted to build a 3D model. It could also be accessible from the program DeepView—an integrated sequence-to-structure workbench. All models are mailed back with a detailed