Review article

Bioinformatics and tools for computer analysis and visualization of macromolecules

Dmitry A. Tikhvinskiy, Yuri B. Porozov

National Research University of Information Technologies, Mechanics and Optics, Saint-Petersburg, Russia

Abstract: The goals and objectives of some topics of structural bioinformatics are presented in the article. The main methods and approaches used in computer biology are highlighted. Areas in which bioinformatic science can greatly facilitate and speed up the work of practical biologist and pharmacologist are revealed. The features of both the basic packages and software devices for complete, thorough analysis of macromolecules and for development and modeling of ligands and binding centers are described.

Keywords: bioinformatics, scientific visualization, macromolecules.

Modern medicine and biology made giant steps in the direction of knowledge of human nature and development of new diagnosis and treatment methods after DNA discovery and publishing «A Structure for Deoxyribose Nucleic Acid» by James Watson and Francis Crick at Nature in 1953 [1]. New research methods such as NMR, immunofluorescence microscopy, PCR (Polymerase Chain Reaction), FRET (Fluorescence Resonance Energy Transfer) have spread everywhere. Techniques of molecular biology allowing to synthesize functionally important proteins for research signaling and metabolic paths of living cells and methods of separation and maintenance of tissue cultures, including blood cells and stem cells are used in researches. Patch clamp method is widely used in electrophysiology, registering current in a small group adjacent ion channels or in a single channel. These methods requires deep understanding of the DNA and RNA structure, the functions of the individual parts and primary, secondary, tertiary and quaternary structure of protein macromolecules, their variability, activity, physical and biological properties.

Bioinformatics (computational biology) is a science, dealing with the nucleic acids in DNA and RNA or amino acids in proteins, their evolution regularities of the above macromolecules, the relationship between the sequence of elements and the spatial structure of macromolecules, its physical properties and functions [2, 3].

According to the European Bioinformatics Institute (EBI), bioinformatics is the use of computer science for biological data management and analysis [4].

This definition implies the use of computers and information technology for the production, storage, analysis and storage of biological data. Bioinformatics is the interface between mathematics, computer science and biology, and the extensive use of mathematical modeling and computational power of desktop computers and multi-cluster systems to meet the challenges of managing large volumes of information.

Main directions of bioinformatics are evolution, search and annotation of genes in a sequence, assembling, annotation and interpretation of genomes, study of exon-intron interactions, gene relationships, classification and characterization of proteins, comparative genomics and proteomics, the evolution of proteins and genomes, phylogeny, structural biology and the development of software packages and network services.

The most important method for analyzing sequence is alignment. There is pairwise and multiple sequence alignment, local and global ones. Alignment is a way of arranging the sequences of DNA, RNA, or protein, based on identity or similarity of subsequences. Evolution causes mutations of genes, insertions or deletions and alignment allow using gaps for proper comparison. This comparison shows nucleotides and amino acids met in each sequence. This conserved sequence regions are often called action sites, they were not changed during evolution. Sequence alignment also can be used to find related proteins or DNAs to follow the evolution of given sequence.

Multiple sequence alignment result analysis helps tracking protein or DNA region's history from first species to modern ones. Using MSA gives information about history of the origin of species. Estimating frequency of mutations gives rough estimate age of species.

Pairwise amino acids alignment on two acids, one of them has unknown tertiary structure, helps to compare their primary structures. Good scoring points indicates homology, similarity of primary structure, functions and 3D-structure of proteins. Proteins that are 40-50% homological will likely have simiral tertiary structures [7]. Functions of this proteins will likely be similar too.
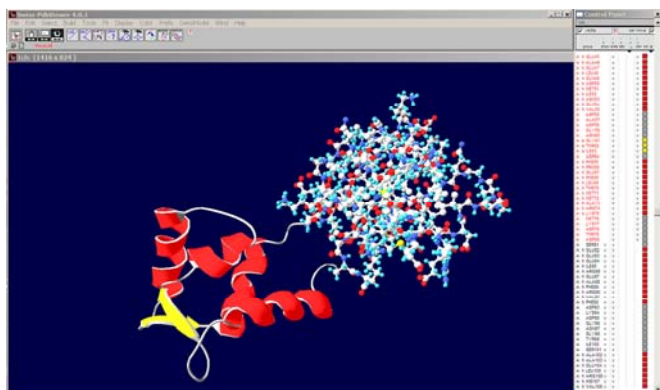
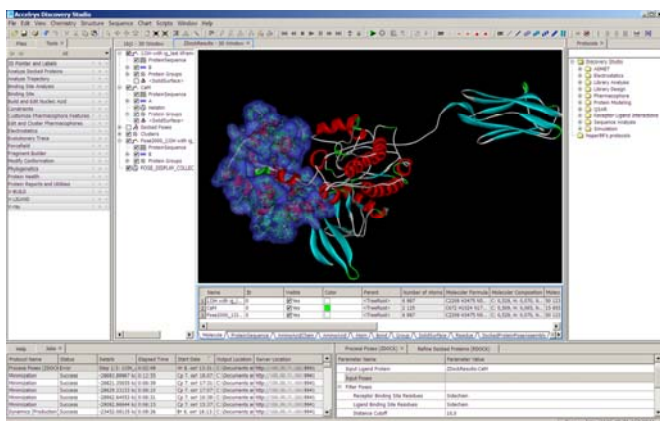**Figure 1. PDB Viewer software displaying 1CFC protein**

Figure 2. Discovery Studio software displaying MLCK (myosine light chain kinase) docking

Protein structural biology [8, 9] is a part of bioinformatics. It widely uses bioinformatics' methods and mathematical tools to solve protein spatial orientation problems. Proved that protein tertiary structure indicates its biological functions, they cannot be identified only by using information about amino acids in protein . X-ray diffraction analysis and NMR are used to determine protein tertiary structure. Results of these studies are stored in a special data banks – PBD, SRS и SRS3d, SCOP, CATH, PFAM etc. Both methods have drawbacks and difficulties of using. Bioinformatics uses different protein tertiary structure prediction and visualization methods [10, 11]: homology modeling, ab initio prediction, statistics-based secondary structure prediction, fold recognition (threading) and (structural alignment). These methods are based on searching homological structures for unknown molecule of using statistically reliable dependencies between certain amino acids sequences and secondary structure elements. This kind of methods cannot determine exact 3D-structure. Ab initio predictions requires enormous computing power. Best results (80-95%) are obtained by using both methods.

Bioinformatics gives information about protein quaternary structure and protein-ligand interaction (docking). Analysing 3D-structures of interacting molecules and their physical properties (hydrophobicity, electrostatic charge, flexibility of individual chains, ability to form hydrogen bonds) allows predict binding sites of these molecules and calculate binding characteristics. This allows modeling of small molecules that can selectively activate or block the active sites of the target protein. This modeling is widely used in development of new drugs, despite the fact that it requires large computational and time resources.

Computational biology also solves the problem of visualization and intuitive representation of biological data. Macromolecules and macromolecule complexes visualization and manipulation instruments required because human receives up to 90% of information from his eyes.

There is special software for operations with protein spatial structures, their groups and protein-DNA, protein-ligand complexes etc. This software allows select structures from data banks and visualize them. Many software packages have tools for creation simple animations, calculating and minimizing molecule potential energy and modeling parts of the molecule de novo, providing tools for in silico mutations, construction loops, reconstruction homologues, changing amino acids rotamers, docking (prediction of binding receptor and ligand), which is extremely important in the design of new drugs. There are free and commercial software packages.

Swiss PDB Viewer (also known as DeepView), developed by the Swiss institute of bioinformatics if free, well-known and powerful software package for protein visualization and modeling [12]. PDB Viewer allows basic operations on the data (*Figure* 1). The package can calculate and minimize the molecule potential energy by the GROMACS96 method, model structures via homology (the procedure is performed on a remote Swiss-Model server), amino acids sequence alignment to build and produce a protein molecules structural alignment. User can perform basic operations on the polypeptide chain: build loops, perform mutations, alter the conformation of the chain using the chart of torsion angles (Ramachandran plot). Script-based integration with PDB viewer is also possible to automate routine operations.

VMD (visual molecular dynamics) and PyMol are more powerful free software. Both support scripting on Python and have good quality of graphics.

There are also more powerful commercial software packages. For instance Accelrys Discovery Studio [13] is a software package, that can solve lots of tasks in molecular modeling. It has well-built UI and advanced graphical engine. Being a complete software package, Discovery Studio can be integrated to Accelrys Pipeline Pilot to model, simulate and construct protein and their complexes, research their interactions dynamically, develop proteins and make QSAR (Quantitative Structure-Activity Relationship). Discovery Studio also allows dock sequences, research protein-binding site properties, run complex AB initio simulations etc. (*Figure* 2). Discovery Studio backend grants access to NCBI (national center for biotechnology information) data banks and instruments, proteomics protocols, pharmacology, sequence analysis etc.

There are special free software for docking too. Autodock and Autodock Vina make it possible to run docking protocols. They are not such user friendly as Discovery Studio but it is possible to automate the process using scripting if one have appropriate experience.

Development of medicine is now directly dependent on molecular level processes understanding. Modern researches in relevant areas as HIV/AIDS and cancer drug research are conducted at the level of genes and proteins controlling transcription and mechanisms regulating these processes. Biology, medicine and IT interactions will become more solid in next decades, so bioinformatics learning at medical and biological universities `is an actual problem.

Reference

1. Watson JD, Crick FHC. A Structure for Deoxyribose Nucleic Acid. *Nature* 1953; 171: 737-8 (PMID: 13054692).

2. Attwood TK, Parry-Smith DJ. Introduction to Bioinformatics. Pearson education ltd, 1999. 238 p.

3. Lesk A. Introduction to Bioinformatics. Oxford, New York: Oxford University Press, 2008. 474 p.

4. The European Bioinformatics Institute [Electronic resource]. URL: http://www.ebi.ac.uk/2can/bioinformatics/bioinf_what_1.html

5. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981; 147(1): 195-7 (PMID: 7265238) (doi: 10.1016/0022-2836(81)90087-5).

6. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 1970; 48(3): 443-53 (doi: 10.1016/0022-2836(70)90057-4) (PMID: 5420325).

7. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. *Embo J* 1986; 4(5): 823-6 (PMCID: PMC1166865) (PMID: 3709526).

8. Branden C, Tooze J. Introduction to Protein Structure. Second edition, New York: Garland pub, 1999. 410 p.

9. Fersht A. Structure and Mechanism in Protein Science: a Guide to Enzyme Catalysis and Protein Folding. New York: W.H. Freeman and Co, 1999. 631 p.

10. Computational Methods for Protein Folding (S.A. Rice and R.A. Friesner eds.); 1st edition. Wiley, John & Sons, Inc., 2001. 544 p.

11. Forster M. Molecular Modelling in Structural Biology. *Micron* 2002; 33: 365-84 (PMID: 11814876) (doi: 10.1016/S0968-4328(01)00035-X).

12. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* 1997; 18(15): 2714-23 (PMID: 9504803).

13. Accelrys Software Inc. [Electronic resource]. URL: http://accelrys.com

Authors:

Dmitry A. Tikhvinskiy – PhD student, School of Natural Sciences, Laboratory of bioinformatics, National Research University of Information Technologies, Mechanics and Optics, Saint-Petersburg, Russia;

Yuri B. Porozov – MD, PhD, Head of the Laboratory of bioinformatics, National Research University of Information Technologies, Mechanics and Optics, Saint-Petersburg, Russia.