PERIODICUM BIOLOGORUM
VOL. 107, No 4, 379–383, 2005

UDC 57:61
CODEN PDBIAD
ISSN 0031-5362

Overview

Why focusing on bioinformatics?

DAVOR JURETIĆ[1]
BONO LUČIĆ[2]
NENAD TRINAJSTIĆ[2]

[1] Faculty of Natural Sciences,
Mathematics and Education
University of Split
N. Tesle 12, 21000 Split
Croatia

[2] Ruđer Bošković Institute, P. O. Box 180
HR-10002 Zagreb, Croatia

Correspondence:
Davor Juretić
Faculty of Natural Sciences,
Mathematics and Education
University of Split
N. Tesle 12, 21000 Split
Croatia

Received November 2, 2005.

Short history

An explosive growth of bioinformatics during last decade tends to hide its much older humble origins as computational biology *(1)*. For instance, the construction of phylogenetic trees *(2)* and the development of protein sequence alignment algorithms *(3)* started almost 40 years ago during late sixties. First secondary structure prediction methods for RNA *(4)* and proteins *(5)* were developed during early seventies. Cracking the »second genetic code«, the protein folding problem, was recognized during seventies as the most fundamental intellectual challenge of computational biology. Major problems attacked already during eighties were sequence analysis, protein structure prediction, molecular evolution, data quality control, collection and free distribution. Under the cover of an older name: computational biology, bioinformatics has matured to become an independent scientific discipline, almost as old as computer science. Central reference data banks and means of accessing them developed in parallel with the extraordinary effort at closing time of last century to decode complete human genome *(6)*.

What is bioinformatics?

Bioinformatics is marriage between computational techniques and molecular biology. It is the application of computational techniques to understand and organize the information associated with biological macromolecules. Its offspring's are many useful algorithms for sorting, retrieving, comparing and analyzing huge amounts of data biologists collected about proteins, nucleic acids and whole genomes. National Institute of Health defined bioinformatics (July 2000, http://www.bisti.nih.gov) as »Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral, or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.« However, bioinformatics is more than just the toolkit collection of algorithms for biologists. Today, bioinformatics is well established discipline in molecular biology, which is uniquely capable with handling a large amount of structural, genomics and gene expression data collected in free-access data-bases, such as GenBank, Protein Data Bank, Swiss-Prot, OMIM, ….

Practicing bioinformatician, like any other molecular biologist, must be capable of posing and answering important questions dealing with structure, evolution and related function of biological macromolecules. Data interpretation is only the last step after data curation and data analysis, which also requires expertise both in informatics and in molecular biology. New research results from the field of bioinformatics are published at ever increasing frequency in a wide spectrum of scientific periodicals usually having high impact factors. This fruitful marriage of informatics and molecular biology can be attributed to the insight that life itself is an information technology.

Relationship of bioinformatics to other life sciences

Bioinformatics connects different areas of molecular biology, genetics and biochemistry. It also connects them with clinical genetics, molecular diagnostics, pharmacogenomics and biomedical informatics. Mathematics, statistics, informatics, physics, chemistry, biology and medicine are all used and connected by bioinformatics. Structural bioinformatics deals with predicting and analyzing structure of biological macromolecules and is closely related to biophysics. Bioinformatics of metabolic processes is related to bioenergetics and to systems biology. Amazing advances in our understanding of evolution would not be possible without bioinformatics.

What are unique characteristics of bioinformatics?

Bioinformatics can put forward several claims of being unique among natural sciences. It offers easily available public databases and software tools that can be used for analysis and decoding of the oldest language, language of genomes and proteomes incised in DNA and protein sequences. It has modest hardware requirements for *in silico* experiments and urgent requirement for young bioinformaticians to devise and perform such experiments. Analysis and synthesis is equally important for bioinformatics. In fact bioinformatics is most productive when used to connect data from different sources. Such holistic approach is obligatory when examining phylogenetic trees, protein-protein interactions and metabolic pathways.

Bioinformatics connects micro and macro world for example in examining how one point mutation in one protein can cause serious genetic disease in the whole organism. It also connects distant past with present time by uncovering surprising evolutionary connections among homologous proteins and genes in organisms as different as yeasts, fruit flies, worms, rats and humans. As a science, bioinformatics is living testimony to the truthfulness of the wise old thought that nothing in the living world makes sense except in the light of evolution.

Service-oriented science

Bioinformatics is also service-oriented science *(7)*. We have been fortunate that vastly improved techniques of raw data collection in molecular biology developed in synergy with advanced software and hardware computer capabilities and with Internet providers bringing this richness to our office or home. Publishing paper in the CC journal with insights from bioinformatics is usually only an obligatory step toward goal of creating Web service that any scientist (or student) can use to gain similar insights concerning his favorite biomacromolecules. Such Web services are one way of increasing individual and collective scientific productivity through e-Science initiative *(8)*. Another public-resource project is based on distributed computing involving huge number of personal computers connected to the Internet. First results for such grid computing projects dealing with the protein folding problem (Human Proteome Folding Project and Folding@Home, http://folding.stanford.edu) have been very encouraging.

In Table 1 several key papers are listed related to protein and genome sequence analysis that are very often used by researchers. These references are sorted according to total number of citations. One can see that excellent (and, very often, simple) bioinformatics solutions have been widely used by molecular biologists, saving time and money for expensive experiments.

Database accuracy

Different genome and protein databases, as public archives, are extremely valuable services that have revolutionized biology. Researchers are regularly using them to hunt for new genes and corresponding protein sequences, to predict protein structure, to discover possible function and to figure out how sequences have evolved in different organisms. It is however naive to assume that the information contained in databases is free of errors and up to date. On the contrary, annotation errors are frequent even in well curated databases such as the SwissProt *(9)*. For any serious research in bioinformatics there is no substitute to extracting accurate information from published papers in addition to data mining from public archives. Bioinformaticist can certainly help biologists to realize not only how much of valuable information is contained in different public databases, but also to understand the limitations connected with their accuracy and with software tools used for analysis. Unfortunately, it is quite easy for biologists to fall into »the black box trap», accepting without question sequence analysis results obtained with chosen software tool, without taking the time to learn technique limitations. Those researchers, who can master limitations and advantages of different biological databases and computational tools needed to exploit them, will have a competitive advantage. On another hand, the grant proposals that lack a bioinformatics component are already being turned down in countries with developed life science research evaluation.

Availability of education in bioinformatics

In anticipation of such development computational biology is now occasionally included as a core module of undergraduate biology. Unfortunately, due to lack of experts instructors for this young science and to difficulties in introducing interdisciplinary subjects in traditional biology curriculum, young biologists are still often educated without obligatory bioinformatics courses. Educating physicists without obligatory mathematics courses, or biophysicists without obligatory biochemistry and molecular biology courses would be equally erroneous.

TABLE 1

The total number of citations (till October 7, 2005) of some basic/well-known bioinformatics papers according to ISI Web of Science database (Philadelphia, USA, http://scientific.thomson.com/products/wos).

Reference	Year of publication	Total no. of citations
ALTSCHUL S F, GISH W, MILLER W, MYERS E W, LIPMAN D J Basic local alignment search tool. *J Mol Biol 215*: 403–410	1990	19404
ALTSCHUL S F, MADDEN T L, SCHAFFER A A, ZHANG J H, ZHANG Z, MILLER W, LIPMAN D J GAPPED BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res 25*: 3389–3402	1997	14879
KYTE J, DOOLITTLE R F A simple method for displaying the hydropathic character of a protein. *J Mol Biol 157*: 105–132	1982	9345
KABSCH W, SANDER C Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*: 2577–2637	1983	3818
CHOU P Y, FASMAN G D Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol 47*: 45–148	1978	2091
HOLM L, SANDER C Protein-structure comparison by alignment of distance matrices. *J Mol Biol 233*: 123–138	1993	1743
ROST B, SANDER C Prediction of protein secondary structure at better than 70-percent accuracy. *J Mol Biol 232*: 584–599	1993	1670
SALI A, BLUNDELL T L Comparative protein modeling by satisfaction of spatial restrains. *J Mol Biol 234*: 779–815	1993	1546
ROST B, SANDER C Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins 19*: 55–72	1994	1046
CHOU P Y, FASMAN G D Prediction of protein conformation. *Biochemistry 13*: 222–245	1974	1041
SANDER C, SCHNEIDER R Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins 9*: 56–68	1991	885
von HEIJNE G Membrane-protein structure prediction – hydrophobicity analysis and the positive-inside rule. *J Mol Biol 225*: 487–494	1992	488
ROST B, CASADIO R, FARISELLI P, SANDER C Transmembrane helices predicted at 95-percent accuracy. *Protein Sci 4*: 521–533	1995	455

Our Ministry of science and National scientific councils in Croatia should support including bioinformatics courses as obligatory both in undergraduate and graduate study of biology, molecular biology, biochemistry, biophysics, medicine...

Predicting membrane protein topology as the case study

It is easy to answer the question why membrane proteins are so important that somebody (like the first author) should spend 15 years of active research life to perfect the algorithm for predicting their topology *(9)*. Integral membrane proteins are coded by 20 to 30% of all genes in both prokaryotic and eukaryotic organisms *(10)*. In the human genome at least 4000 genes are coding for integral membrane proteins critically important for signal transduction, nerve conduction, cell-cell interactions, solute and macromolecular transport across membranes, ATP synthesis, hearing, vision, olfaction and pain detection. Membrane proteins are targets of choice for most drugs. However, due to their hydrophobic nature membrane proteins are difficult subjects for isolation and for structural studies. As a rule they do not crystallize and are hardly tractable by NMR spectroscopy. They account for less than 1% of the proteins with determined X-ray structure *(11)*. In the absence of high resolution three-dimensional structures, uncovering structure-function relationships for membrane proteins depends on availability of accurate topological models for the number, orientation and sequence location of transmembrane spans (usually in the α-helix conformation). Web servers for automatic sequence analysis and prediction of transmembrane topology have proliferated during recent years so that it became necessary to perform objective performance tests of their accuracy *(12)*. The topology predictor SPLIT developed at the University of Split, (http://split.pmfst.hr/split/4/) and at the University of Osijek, Croatia, is one of the most accurate predictors. This is somewhat surprising result, because SPLIT does not use the multiple alignment information reported to increase prediction accuracy *(13)*. In this very competitive field our predictor has good potential to remain top performer if following improvements are incorporated:

a. Prediction of hydrophobic signal sequences will decrease false positive predictions when signal sequences are mistakenly predicted as a transmembrane helix.

b. Multiple alignments of homologous sequences to the tested sequence will use rich evolutionary information to increase prediction accuracy.

Presently, SPLIT is used about 800 times per month for sequence analysis of integral membrane proteins. Almost all leading universities from USA, Europe, China and Japan have been using it during years 2004 and 2005.

Some important insights from bioinformatics have not been expected

As The Three Princes of Serendip *(14),* Science enriches our lives by making desirable but unsought-for discoveries. It does not matter in the end what was the original motivation and research goal. All good scientists have the faculty of serendipity. That is why investing in science, as a long term commitment, has been wise strategy for countries that adopted it. We shall mention only two of many recent examples when bioinformatics produced unsought-for discoveries.

After genomes of several hundred organisms have been sequenced it becomes clear that even genes of species that separated several hundred million years ago are still similar. For instance, better insights into some human genetic diseases have been achieved by exploring related yeast genes. In fact sequence analysis revealed that 50% of human genes have homologues in yeast, fruitflies or worms *(6)*.

Horizontal gene transfer among unrelated species is much more frequent than previously believed. Even the origin of the eukaryotic cell might have resulted from a fusion of eubacterial and archaebacterial prokaryotic genomes *(15)*. This novel insight from bioinformatics is in essence the proposal to replace Charles Darwins description of the phylogenetic tree of life with a ring of life description. A ring of life implies promiscuous gene sharing leading to the last universal common ancestor (LUCA) of all eukaryotic cells.

Some challenges for the future

Our exalted opinion of ourselves with respect to all other life forms is still trying to come to terms with recent revelation from bioinformatics *(16)* that our genetic wisdom expressed through 20000 to 25000 genes is barely better than that of the worm *Caenorhabitis elegans*. A tiny puffer fish *Tetraodon nigroviridis* commonly kept in aquaria, with the smallest known verterbrate genome, is also thought to have between 20000–25000 protein coding genes *(17)*. We do not know why do humans have so few genes. We still have very limited knowledge concerning the question what coding or noncoding DNA regions are responsible for accelerated evolution of human species and how the activity of small total number of protein-coding human genes is controlled during development. It seems that RNA world is still alive and well, as revealed through the abundance of DNA codes for short RNA *(18)*. Important regulatory function has been discovered for some of these codes, but largely their function is still unknown. Comparative genomics uncovered that far from coding genes many conserved non-coding sequences exist with unknown function *(19)*. It is the challenge for the bioinformatics to figure out such unknown and no-doubt important functions for genome regions previously considered as junk DNA.

While DNA is very stable molecule, genome and proteome is much more flexible, subject to strong opposing forces of mutations and natural selection. Man-made environment is becoming ever more important in directing evolution of genome in many species including our own. Much more research in bioinformatics is needed before we can gain clear insight where the experiments of global civilization are leading us in terms of genome and proteome evolution.

Protein folding problem is still huge challenge, which probably requires novel ideas from physics, chemistry and informatics to speed up simulated evolution of unfolded into folded protein state. Bioinformatics can help to test each new idea against real structures that biological evolution developed through the trial and error method.

REFERENCES

1. OUZOUNIS C A, VALENCIA A 2003 Early bioinformatics: the birth of a discipline – a personal view. *Bioinformatics 19*: 2176–2190
2. FITCH W M, MARGOLIASH E 1967 Construction of phylogenetic trees. *Science 155*: 279–284
3. CANTOR C R 1968 The occurrence of gaps in protein sequences. *Biochem Biophys Res Comm 31*: 410–416
4. TINOCO I, UHLENBECK O C, LEVINE M D 1971 Estimation of secondary structure in ribonucleic acids. *Nature 230*: 362–367
5. CHOU P Y, FASMAN G D 1974 Prediction of protein conformation. *Biochemistry 13*: 222–245
6. LANDER E S *et al.* 2001 Initial sequencing and analysis of the human genome. *Nature 409*: 860–921
7. FOSTER I 2005 Service-oriented science. *Science 308*: 814–817
8. HEY T, TREFETHEN A E 2005 Cyberinfrastructure for e-science. *Science 308*: 817–821
9. JURETIĆ D, ZORANIĆ L, ZUCIĆ D 2002 Basic charge clusters and predictions of membrane protein topology. *J Chem Inf Comput Sci 42*: 620–632
10. KROGH A, LARSON B, VON HEIJNE G, SONNHAMMER E 2001 Predicting transmembrane protein topology with a hidden Markov model. Application to complete genomes. *J Mol Biol 305*: 567–580
11. WHITE S H 2004 The progress of membrane protein structure determination. *Protein Sci 13*: 1948–1949
12. KERNYTSKY A, ROST B 2003 Static benchmarking of membrane helix predictions. *Nucl Acids Res 31*: 3642–3644
13. PERSSON B, ARGOS P 1996 Topology prediction of membrane proteins. *Prot Sci 5*: 363–371
14. The man in question was **Horace Walpole** (1717–97), fourth Earl of Orford, son of Prime Minister Robert Walpole, connoisseur, antiquarian and author of the famous gothic novel, *The Castle of Otranto* (London, 1765). The word he invented was, of course, *serendipity.* And the tale he rescued from literary oblivion was *The Three Princes of Serendip*. The letter – to Horace Mann, an envoy in the service of King George II stationed in Florence – was written to acknowledge

the safe arrival of a portrait of Bianco Capello, a 16th century beauty and Duchess of Tuscany. This letter is contained among the 31 volumes of *Horace Walpole's Correspondence* (New Haven, 1937), edited by Wilmarth Sheldon Lewis. (taken from http://livingheritage.org/three_princes.htm)

15. RIVERA M C, LAKE J A 2004 Ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature 431:* 152–155

16. International Human Genome Sequencing Consortium 2004 Finishing the euchromatic sequence of the human genome. *Nature 431*: 931–945

17. JAILLON O *et al.* 2004 Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature 431*: 946–957

18. CHENG J *et al.* 2005 Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science 308*: 1149–1154

19. HILLIER L W *et al.* 2004 Sequence and comparative analysis of the chicken genome provide unique perspectives on verterbrate evolution. *Nature 432*: 695–716.