BioTechnologia vol. 92(3) • pp. 217-220 • 2011
Journal of Biotechnology, Computational Biology and Bionanotechnology EDITORIAL

On the border between biology, mathematics and computer science

PIOTR FORMANOWICZ
Institute of Computing Science, Poznan University of Technology, Poznań, Poland
Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznań, Poland
e-mail: piotr@cs.put.poznan.pl

Computational biology, bioinformatics and systems biology are three closely related interdisciplinary areas of research rapidly evolving at the intersection of biology, mathematics and computer science. The general goal of all of these areas is to support the analysis of biological phenomena by mathematical and algorithmic methods. However, the ideas flow not only from mathematics and computer science to biology but also in the opposite direction when the biological objects and processes become a motivation for developing new mathematical theories or algorithms. It may be said that, at least in some sense, bioinformatics and systems biology are parts of computational biology. While computational biology as a whole, concerns the development of mathematical models of biological phenomena and algorithmic methods and tools for the analysis of biological processes and structures, bioinformatics is focused mainly on using these tools for biological discoveries. On the other hand, systems biology is related to the systems analysis of biological complex objects which, in turn, includes the development of mathematical and algorithmic methods for such an analysis and using them for biological investigations.

For a long time, biology, on one hand, and mathematics and computer science, on the other, have been seen as almost distinct areas of research. It was mostly due to the fact that in biological sciences, as opposed to, for example, physics, qualitative (but not quantitative) descriptions and analyses were the basis for scientific investigations. This situation started to change when it became possible to read and analyze the most important building blocks of living organisms, i.e. nucleic acids and proteins.

The impressive discoveries in molecular biology during the second half of the 20th century and especially the development of new very effective DNA sequencing techniques and methods at the beginning of the new millennium have resulted in an enormous amount of new biological data available in publicly accessible databases. Not only it has accelerated research in almost every branch of biology but has also demonstrated that the analysis of these data is very difficult or even impossible without mathematical and algorithmic methods. The reason for this is twofold. Firstly, the analysis of large sets of data, in general, requires formal methods. Secondly, it seems that the structure and functionality of living organisms is based on some strict rules which, at least in principle, can be described using a language of an appropriate mathematical theory. Previously, this was not so obvious but now there are strong evidences that this is the case.

First of all, the entire organization of genetic information in genomes is very complex; however, at the very basic level, it is also very regular. The genetic code is a simple but rigorous and effective method which the nature applies to store and process the genetic information. Although it is simple, its discovery was not a trivial task. Genetic information is written in DNA and RNA molecules as sequences of nucleotides and therefore it can be represented in a natural way as sequences of letters corresponding to these nucleotides. Objects of this type (i.e. sequences of letters) are well known in computer science which has made it a subject of an intensive research from its very beginning. This makes many molecular biology issues well suited for the analysis using computer science–based methods. But more important and fundamental is the fact that living organisms are based on information processing. Indeed, the genetic information must be stored, copied and corrected, and on the basis of this, the structure of an organism is built and its functionality determined. Information processing is the core of research in computer science and related disciplines. Although until recently computer science was focused on information processing in technical systems, the general idea in both, the technical and the biological ones, is the same. However, biological systems seem to be more challenging since they are more complex and their exact structure is usually unknown.

At this point, it should be noted that it is only recently that biological phenomena have been treated as complex systems. As a result of the above-mentioned accumulation of the huge amount of available biological data, it is becoming more and more obvious that the structure and the functionality of living organisms follows not only directly from the nature of their basic building blocks but also from the interactions between these blocks. So, it seems that the dominant and very successful approach in biological sciences focused mostly on the precise analysis of basic elements of living organisms has serious limitations since the inherent complexity of life, most probably, lies in the dense nets of interconnections between these elements. This is the reason that not only living organisms, but also their parts like organs, tissues etc., should be treated as systems. Here, again we turn to mathematics, computer science and related disciplines, where systems have been analyzed for years, mostly in the context of technical sciences. One of the most important features of the system-based research is the fact that, at least to some extent, systems can be studied at a very general level as mathematical objects and the theory developed in this way can be further applied to the description and analysis of particular physical, technical or biological systems. So, when it was realized that biological objects are in fact complex systems, it became possible to apply known systems theories in the area of biological research. Obviously, using only the methods developed previously in order to study, for example, technical systems, may be insufficient since the biological systems are, in general, the most complex systems analyzed so far. Hence, developing new and more advanced systems theory is necessary.

It should also be mentioned that in the case of biological systems, even the development of their precise formal models is in most cases a challenging task. It follows from the obvious fact that biological systems are developed by the nature but not by humans so their exact structure has to be discovered. The process of constructing a precise model of a biological system is in many cases closely related to discovering the structure of the analyzed biological phenomenon since often in order to build an exact model some open questions concerning this phenomenon should be answered.

The systems approach in biological sciences seems to be very promising because, at least in principle, it can capture the complexity of living organisms, which is, most probably, one of the fundamental features of life. Moreover, with technical systems as a reference, one of the distant goals of this approach is (or at least it should be) to develop precise methods for controlling biological systems. This means that on the basis of a very precise model, it should be possible to control the behavior of a biological system by applying external stimuli in a manner that follows from the analysis of the model. Achieving this goal would be extremely important for, among others, medicine since a disease can be seen as an abnormal state of the biological system, which in this case is the human body. So, a precise method of changing this state to a normal one would be equivalent to an effective therapy of the disease. But there is a long way to achieving this goal, because constructing even an approximate model of a biological system of moderate size and complexity is usually a hard task.

The branch of computational biology involved in systems approach to the analysis of biological phenomena is called systems biology and is now an area of very intensive research. It is interesting that their roots can be found in the mid 1940s when cybernetics was discovered by Norbert Wiener or even earlier, in the 1920s when Ludwig von Bertalanffy postulated the need for a theory concerning systems in general, which is now called general systems theory. It is interesting that the inspiration in both of these cases were the biological phenomena, even though the theories then evolved in the direction of technical systems. So now, at least in some sense, the return to the roots of systems sciences can be observed.

Although at this moment, systems biology seems to be the most important part of computational biology, many of its other branches are also fundamental for many areas of biological research. The classical branch of computational biology is sequence analysis. As has been mentioned before, sequences are natural objects for computer science methods, so in the early stages of computational biology history problems directly related to sequences were most intensively studied. Here, at least four main classes of problems can be distinguished, i.e. sequence comparison, searching for motifs, phylogenetic analysis and reading DNA sequences.

Comparing two or more nucleotide or amino acid sequences may lead to a discovery of some similarities which, for example, may be a result of similar functions performed by genes or proteins whose sequences are being compared. Classical computational methods for sequence comparison are based on dynamic programming. Problems of this type attracted mathematicians and computer scientists a long time ago and one of them was Stanisław Ulam who is considered to be one of the founders of computational biology. Closely related to sequence comparison is searching for motifs. While the general goal in the case of sequence comparison is to determine the similarity of the analyzed sequences to each other, here the goal is to find some subsequences which have some common features, i.e. which can be described using a certain pattern. The results of sequence comparison can be a starting point for searching for motifs, although methods of different kinds are used for this purpose in general.

Another very classical branch of computational biology is a phylogenetic analysis. The main task of this area is to construct a phylogenetic tree representing a hypothetical evolutionary history of a group of species. The tree is constructed on the basis of similarities between these species and from the computer science point of view it leads to very interesting and challenging problems. As it is easy to guess, nowadays the concrete basis for such a construction are usually nucleotide or amino acid sequences of a gene or protein present in all of the analyzed species. This area is also one of the oldest branches of computational biology since phylogenetic trees were constructed even before the molecular biology era. In those times no formal methods were used for building such trees. However, since, on the one hand, they are graphs (i.e. mathematical objects well-known in discrete mathematics) and on the other hand, similarities in input data being the basis for reconstruction of evolutionary history can be analyzed using formal mathematical and algorithmic methods, phylogenetic analysis has become an area where very advanced methods of this type are used. However, since various algorithms can provide different results on the basis of the same input data, the algorithm for phylogenetic tree reconstruction should be carefully chosen, taking into account the specificity of the analyzed biological process.

One of the most important areas of computational biology was for a long time reading DNA sequences. Historically, the process of reading DNA sequences was divided into three stages, i.e. sequencing, assembling and mapping but nowadays all of them (at least to some extent) are included in modern sequencing methods. Obviously, in order to perform any analysis of nucleotide sequences first they have to be read, so DNA sequencing is in fact the first and the necessary step in the process of genetic information analysis. In general, computational problems arising here concern determining the proper order of short DNA fragments obtained in the biochemical phase of a sequencing procedure. In practice, problems of this type are difficult because of errors occurring in the biochemical stage. The rapid development of next generation sequencing techniques is the reason for active research in the area of sequencing algorithms.

Another, very important and broad areas of research are methods for the analysis of microarray data. Microarrays are a tool mainly used for gene expression analysis which, in principle, allows for precise investigation of functions of many genes in parallel. Although the technology used here is well-developed, the bottleneck is the analysis of data coming from microarrays. Here, mainly statistical based methods are used and a broad spectrum of those is available. The problem lies in answering the question which method should be used to analyze a particular set of data. This question is extremely difficult because the method should be carefully chosen to fit the nature of the biological process being the source of the data. It is a very hard task that requires deep understanding of both, the data analysis methods and the biology. Answering this question properly is the basis for reasonable microarray analysis since various methods can provide completely different results on the basis of the same data, which may lead to wrong biological conclusions.

One of the classical branches of computational biology is the prediction of protein 3D structures. Since spatial structure of the protein is crucial for its functionality, methods for determining such structures on the basis of a sequence of amino acids are intensively investigated. But it is a kind of the Holy Grail of computational and molecular biology – no satisfactory algorithm is known and strong competition in this area between research groups all over the world is taking place. Analogous problems concern the prediction of 3D RNA structures; however they seem to be a little easier.

The above mentioned branches of computational biology are only some "classical" examples and it is worth mentioning that the results obtained in all those areas are nowadays very often elements of systems analysis of biological processes.

Obviously, since biological sciences evolve very rapidly new challenges for computational biologists constantly appear and new kinds of mathematical and computer science problems inspired by biology must be formulated and solved. In an ideal situation a close collaboration between biologists, on one side, and mathematicians and computer scientists, on the other, should take place during the whole process of investigating complex biological problems. Without such a collaboration usually it is very difficult, or even impossible, to thoroughly explore the biological reality using formal methods and tools. In such a case, on the one hand, the mathematical and algorithmic problems inspired by biological phenomena are usually not very exciting for mathematicians and computer scientists, and on the other hand, the biological discoveries made on the basis of the solution of these problems are not very spectacular (although, they can be interesting). But when such a collaboration is possible, the formal problems defined on the basis of a detailed observation of biological objects and processes can be interesting for mathematicians and computer scientists since they can open some new areas of research on mathematical theories or advanced algorithms and the theoretical results obtained using them can be applied to make crucial biological discoveries as they can capture some hidden features of the studied biological phenomena.

Despite that, using mathematical and computer science methods and tools for investigating biological reality is so fruitful it is worth mentioning that one should be careful while drawing biological conclusions on the basis of the results provided by bioinformatics tools. The reason is that each such tool (i.e. a computer program) is an implementation of some algorithm which is based on a mathematical model of the analyzed biological phenomenon. It is a source of errors of at least two types. First, each model is only an approximate description of some fragment of physical reality. This approximation is the source of errors of the first type. To be more precise, the model is often a mathematical problem which should be solved. The solution corresponds to the solution of the analyzed biological problem but need not to be equivalent to it. In other words, the exact solution of the mathematical problem is usually only an approximate solution of the biological one. Hence, it is a source of the already mentioned errors of the first type. Moreover, since the mathematical problem is often computationally intractable (i.e. NP-hard), the algorithms used for solving it are heuristics (because in such a case, exact algorithms are usually inefficient). So, they solve the mathematical problem in an approximate way which is a source of the errors of the second type. From this it follows that usually when using a bioinformatics tool for solving some biological problem one gets an approximate solution of some mathematical problem which is only an approximate description of the studied biological phenomenon (problem). This is the reason why one should be very careful in drawing biological conclusions on the basis of such solutions.

This is also the reason for developing more accurate models and algorithms, because the more precise they are, the better justified the biological conclusions will be.