Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Opinion

Bioinformatics: perspectives for the future

Luciano da Fontoura Costa

Cybernetic Vision Research Group, Institute of Physics at São Carlos, University of São Paulo, Caixa Postal 369,
13560-970 São Carlos, SP, Brazil
Corresponding author: L. da F. Costa
E-mail: luciano@if.sc.usp.br

Genet. Mol. Res. 3 (4): 564-574 (2004)
Received October 4, 2004
Accepted December 10, 2004
Published December 30, 2004

ABSTRACT. I give here a very personal perspective of Bioinformatics and its future, starting by discussing the origin of the term (and area) of bioinformatics and proceeding by trying to foresee the development of related issues, including pattern recognition/data mining, the need to re-integrate biology, the potential of complex networks as a powerful and flexible framework for bioinformatics and the interplay between bio- and neuroinformatics. Human resource formation and market perspective are also addressed. Given the complexity and vastness of these issues and concepts, as well as the limited size of a scientific article and finite patience of the reader, these perspectives are surely incomplete and biased. However, it is expected that some of the questions and trends that are identified will motivate discussions during the IcoBiCoBi round table (with the same name as this article) and perhaps provide a more ample perspective among the participants of that conference and the readers of this text.

Key words: Bioinformatics, Post-genomics, Modeling, Computing, Physics

INTRODUCTION

Science is one of the principal manifestations of human curiosity, and so far its best means to understand, control and predict nature. As such, scientific endeavor is developed *by* humans *for* humans, meaning that the results should ultimately satisfy not only our needs, but should also be expressed in a way that is accessible to our very particular ways of looking and interacting with our world. Examples of human bias include the predominant importance given to stimuli in the visible spectrum and the sequential nature of our thoughts. While the human perspective is itself changed by science, to the extent that we are now able to develop highly abstract and sophisticated models (science can be understood as the art of model building), we remain a very particular beginning and end of all science (Costa, 2003).

Created (or perhaps just discovered) by humans, mathematics became the basic language of science, providing an objective and impersonal medium for organizing, in a formal way, the concepts and relationships that are involved. As the first science to adopt the mathematical approach in a systematic way, physics became the reference for other areas as they progressed towards formalization through mathematics. After the mathematization of chemistry, an endeavor inherently based on quantum mechanics, biology stands as the next candidate. And what a candidate! While a great part of the success of physics, at least at its more initial stages, is a consequence of its reductionist approach, the highly complex web of biological connections, extending along wide scales of space and time, together with its historical and statistical nature, imply new creative challenges and approaches. Such challenges can only be met by the comprehensive use of modern informatics, hence the new area of *Bioinformatics*. At the same time, the huge rewards awaiting the taming of the biological world, including longer lives and the cure of diseases, have been a major additional drive to biological research, with consequences for human resource formation and industry, along with ethics, morals, and even religion.

After making a review of some basic issues underlying the genesis of bioinformatics, I saw that there is a need to see what the future holds for this emergent area, including the challenges implied by pattern recognition, the reintegration of biology, the potential of complex networks, the exciting perspective of uniting neurobioinformatics, how to form human resources, research and industrial perspectives, and possible implications for ethics and morals.

HOW DID *INFORMATICS* GET INTO *BIOINFORMATICS*?

Although there is little controversy about the meaning of *Biology* (e.g., "the study of living beings") and *Informatics* (e.g., "the science of acquiring and manipulating data"), a lot of misunderstanding and disagreement can be sparked by the simple juxtaposition of these two terms in order to obtain *Bioinformatics*. The main reason for such complications is that this word does not usually mean its obvious interpretation as the *use of informatics to study living beings*. Rather, bioinformatics has, for historical reasons, been closely associated with the application of informatics to the study of *genetics* data. Perhaps this narrower interpretation was motivated by the lack of a suitable word to express the latter idea. Indeed, alternatives such as *geneinformatics*, *geninformatics*, *DNAinformatics*, all sound rather awkward. By being the first to embrace informatics in a systematic fashion, genetics had the opportunity to choose the simplest name, i.e., *bioinformatics*, while latecomers such as neuroscience, immunology and

phylogenetics had to content themselves with more restrictive (and proper) names such as *neuroinformatics*, *immunoinformatics*, and *phyloinformatics*.

While the author of these lines has, rather unfortunately, no reasonable alternative word to express the application of informatics to genetics, he believes in the broader and more logical understanding of *bioinformatics as the application of informatics to biology* (see Figure 1 for terminology). The immediate advantage of proceeding in this manner is that it emphasizes the fact that biology is, ultimately, an integrated realm that cannot be comprehensively understood by reductionist approaches (see the Section Reintegrating Nature: Systems Biology).

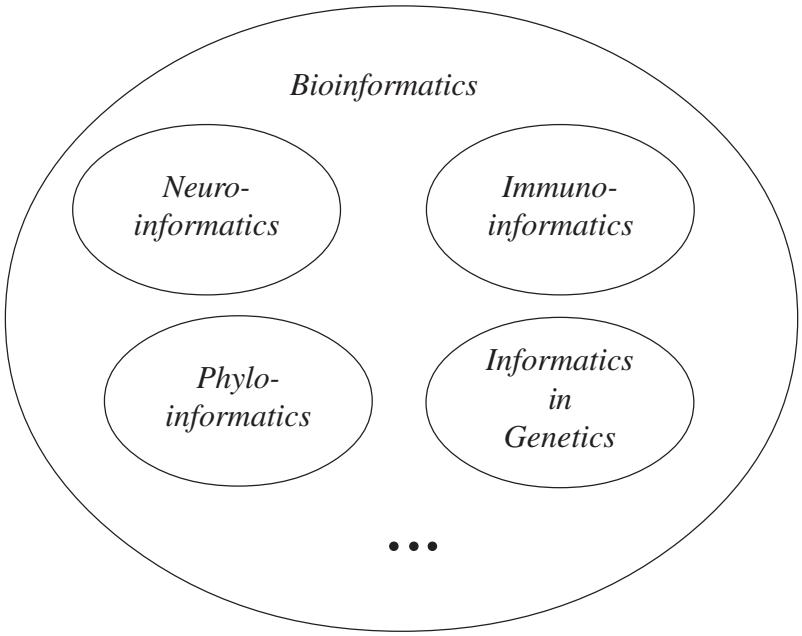

Figure 1. The *bioinformatics Easter Egg*: bioinformatics as the general area involving the application of informatics to biology. Some possible derived subareas are also represented.

An important point to be recalled is that by informatics, we mean not only the traditional (and important) use of computers for numerical methods, which characterizes the area called *computational biology*, but also the use of *modern concepts* from computer science, including:

Databases

Databases are required to organize and access in an effective way the vast amount of data characterizing most biological problems. While this area is relatively well developed, issues such as content-based retrieval, especially those based on visual properties, represent several scientific-technological challenges.

Internetworking

The distribution and sharing of data and methods is essential not only for distributing the

computational demands implied by the large databases, but also to cater to effective interaction between members of multidisciplinary teams. Networking can be performed locally (e.g., LANs) and on a wide scale (Internet).

Parallel computing

The large amount of data, allied to the complex nature of biological interactions, often imply the use of parallel/concurrent processing machines and systems. Special interest has been focused on clusters of personal computers, which have been extensively used for biological analysis and simulation, while grid computing remains an interesting perspective for larger-scale integration and number crunching.

Image analysis

The scientific-technological advances in data acquisition have allowed the design and implementation of acquisition devices and instruments capable of producing 2D, 3D and even 4D (three spatial coordinates plus time) representations of biological structures and phenomena. A promising perspective is the application of concepts and tools from image analysis to the characterization, classification, modeling, and simulation of such data, including spatiotemporal profiles of gene expression in developing organisms.

Artificial intelligence

Artificial intelligence can be used to automate the analysis and identification of rules underlying biological data. Particular relevant sub-areas of artificial intelligence are pattern recognition and datamining (see the Section Datamining versus Modeling).

The rules played by these areas in bioinformatics become clearer as we progress further into this question. We have made highly simplified diagrams of mathematical model construction (Figure 2) before (A) and after (B) the introduction of computing/informatics. The natural phenomenon of interest is represented by a limited number of variables and equations, defining a possible model to be validated by comparing (the block marked as "C") its predictions experimentally. The introduction of computing allowed the models to be systematically simulated in computational fashion. The incorporation of modern informatics (dashed box) concepts and tools pervades the whole approach (B), presenting potential for enhancing and automating almost all involved tasks.

THE FUTURE OF BIOINFORMATICS

The following subsections present a series of important issues related to or implied by bioinformatics, starting from the scientific investigation perspective and moving into human resource formation and industrial aspects.

Datamining versus modeling

One of the main activities underlying bioinformatics is *datamining*, namely the use of

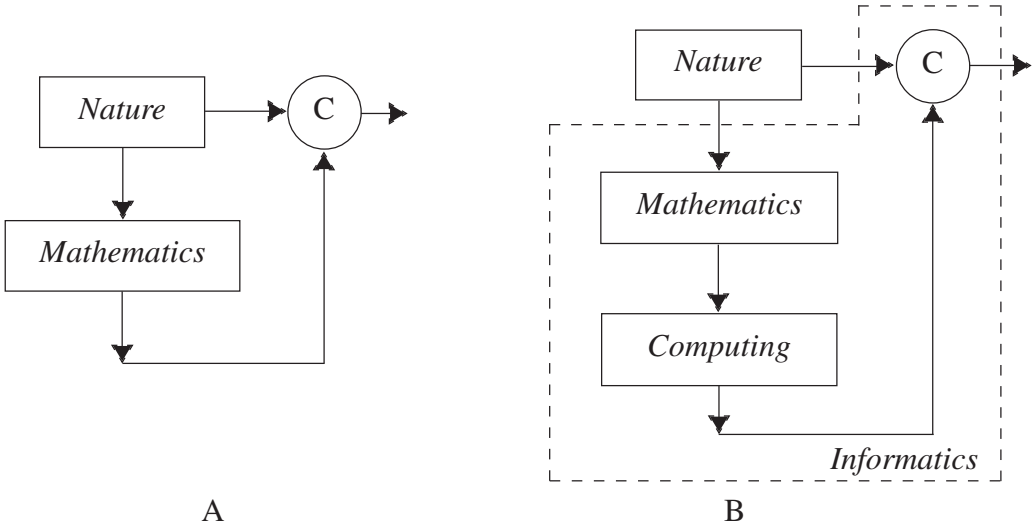

Figure 2. The formal approach to science before (A) and after (B) computing/informatics.

computers to seek automatically for patterns, structures and rules in large databases. As such, datamining can be understood as an extension of the area of *pattern recognition*, where one studies methods for identifying patterns in not necessarily so large databases. The essential problem of pattern recognition can be easily understood from the three simple diagrams in Figure 3.

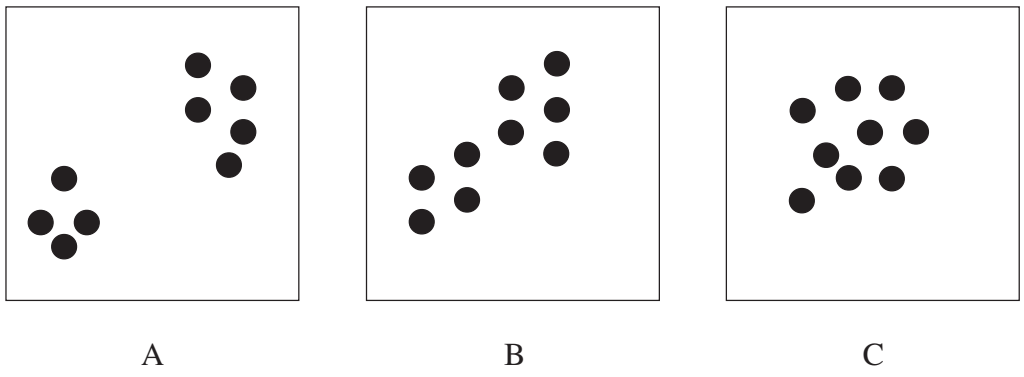

Figure 3. The essence of pattern recognition: while two classes are clearly perceived in A and a single class is seen in C, there is no means for a human to tell if B corresponds to one or two classes.

While two groups are immediately perceived in Figure 3A and a single group is seen in Figure 3C, it is rather difficult to conclude whether the dots in Figure 3B belong to one or two groups (you will very likely get different answers from different people). As the science of trying to find an answer to such an issue, pattern recognition stands out as one of the most difficult and subjective scientific activities. Although elegant mathematic and statistical con-

cepts and methods have been developed and applied to cope with this important problem (see, for instance, Duda et al., 2000; Costa and Cesar, 2001), the subjective nature in which problems are usually formulated - combined with a number of other issues, such as statistical sampling and the decision to use normalized or dimensional data, often conspire to undermine the whole enterprise, or at the very least make it a highly subjective. The important point to be borne in mind is that, unless some objective mathematical merit function is supplied, the solution of the difficult case in Figure 3B will ultimately depend on human judgment (e.g., Gestalt), which is inherently subjective. This point is especially important because very often the human inspection of the obtained measurement spaces, achieved through scientific visualization, completely defines the conclusions regarding the structure of the clusters in the data.

As a successor to pattern recognition, datamining often suffers from the same inherent problems of subjectivity, now amplified by the large size of the database and the more automated nature of the methods adopted for the mining.

Reintegrating nature: systems biology

While biology is a single connected whole, scientific approaches to understanding biological phenomena have often relied on *reductionism*. Such an approach is characterized by the focus of interest on some particular part of the phenomenon of interest, which is isolated from the rest of nature as thoroughly as possible. Such a divisive and isolationist approach is interesting and effective because, by limiting the effect of other elements on the problem of interest, it enhances our chances of understanding the problem and obtaining a suitable model. Although the reductionist approach applied to biological systems has provided a wealth of findings and knowledge about nature, it is reaching the rock bottom of specialization. In other words, to continue learning about a specific biological system it is necessary to look around and make connections with other systems and structures. After all, life is a direct consequence of a long, dynamic, evolutionary process, involving a very wide range of spatial and temporal scales. How can we understand the beak of a given species of hummingbird without considering the flowers from which that small creature feeds? And, to understand the flowers, we need to go further and consider the environment where they developed. Ultimately, it is impossible to isolate life from the rest of our universe in the medium and long term. As illustrated in Figure 4, we are currently involved in a process that will lead to the eventual reunification of biology and nature.

Complex networks: the key to science unification?

The new area of complex networks (Albert and Barabási, 2001; Newman, 2003), which can be understood as the integration of graph theory and statistical physics, is poised to become one of the most interesting research areas in this century.

The origin of the area of complex networks can be traced back to the pioneering works of Erdös and Rényi (1959) on *random networks*. Such discrete structures can be obtained as follows: given N initially isolated nodes, links are added at random between any two nodes. Interestingly, as the number n of edges increases, the nodes become more and more connected, and trees and cycles start to appear. At a critical edge density, the network undergoes a major topological transformation, in the sense that a giant cluster appears. Such an abrupt change of the properties of the random network is a nice example of the concept of *phase transition*,

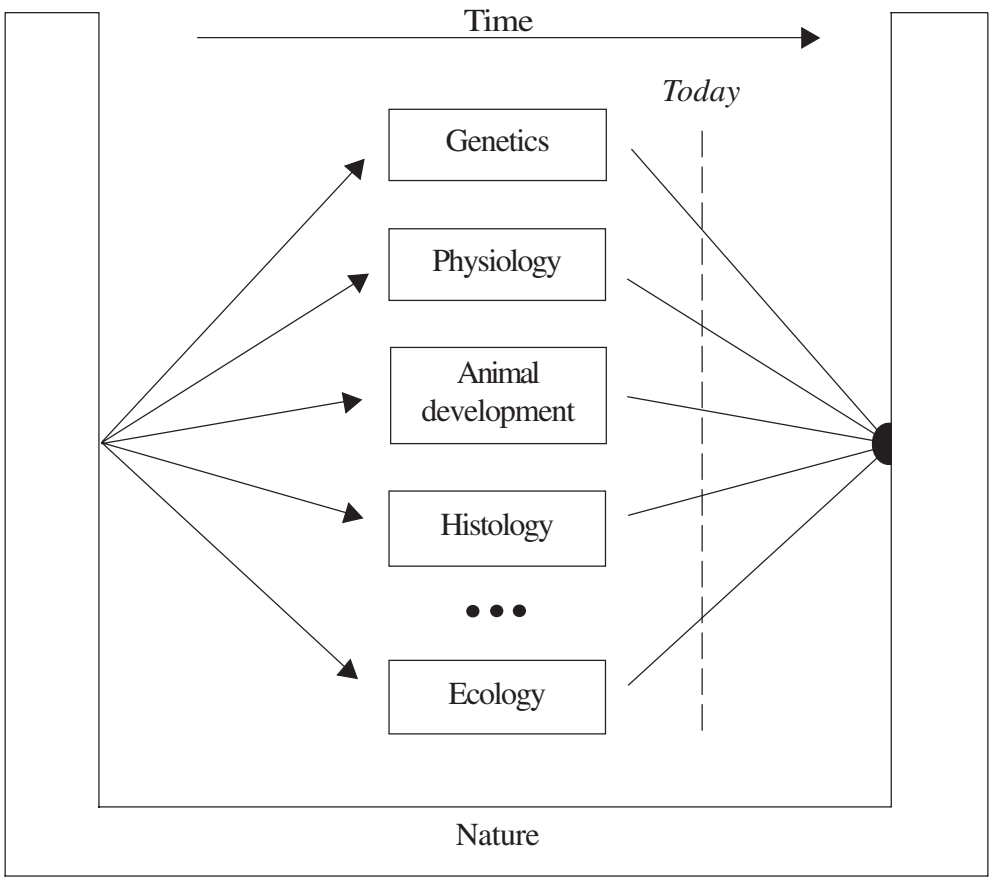

Figure 4. The division (specialization) and reintegration of biological studies. We live at a time close to that marked by the dashed line.

more specifically a *percolation*. Although Erdös and Rényi accomplished a very comprehensive mathematical study of random networks, such structures were unfortunately found not to be good models of natural structures or phenomena. Later, developments by Watts and Strogatz (1998) and other scientists from various areas targeted the interesting type of networks characterized by a relatively small average number of edges between any two nodes, which were duly called *small world networks*. While such networks were successfully used to model social relationships, the small world property turned out to be rather general, being also characteristic of random and many other types of networks. Recent developments by Barabási and other scientists led to the interesting network model known as *scale-free networks*. Such structures are characterized by the fact that they do not have a typical *node degree*, which is formally defined as the number of edges attached to a node. Indeed, dilog plots of the distribution of node degree are found to converge to straight lines, without any typical trend. Such a structural property has the important implication that nodes with a high degree - the so-called *hubs* - become more likely to appear. Hubs are all important because they dominate the connectivity of the network. For instance, an attack on a few hubs will quickly dismantle a scale-free network (and also many other types of networks).

Complex networks have been applied to several important problems in bioinformatics, with encouraging success (e.g., Jeong et al., 2000, 2001; Wagner and Fell, 2001; Bose, 2002; Holme et al., 2003; Vazquez et al., 2003; Costa 2003a). For instance, it has been shown that protein-protein interaction networks typically follow a scale-free topology, and that essential proteins tend to correspond to hubs. Promising results have also been obtained for modeling and characterizing gene activation networks.

Although a great part of the initial attention given to complex networks was motivated by the fact that several important systems, such as the Internet, tended to follow scale-free topologies, interest has recently been focused on a series of complementary investigations, such as the identification of *communities* in networks (i.e., groups of more intensely connected nodes), the study of the *dynamics* of processes defined over the network topology (e.g., systems of differential equations where each node corresponds to a variable and the edges represent the coupling between such variables), and the extension to all types of networks, including those in which the nodes are allowed to move in the spatial domain. Special attention has also been drawn to generalizing the concepts and measurements of complex networks, including the possibility to define node degree and clustering coefficient to subsets of the networks (Costa, 2004). Such trends, allied to the inherent potential of graphs for representing virtually any discrete structure (including trees, vectors, lists, queues, and so on), as well as the possibility to run the quite varied dynamics over such structures, makes complex networks a primary candidate for integrating several scientific areas, and perhaps providing a basic representational framework leading to the unification of science.

Neurobioinformatics

While there is more to biology than genomes, the remarkable scientific-technological advances in this area along the current decade have implied that most biological investigations are now expected to incorporate the genetics component, i.e., to be viewed also from the gene perspective. Indeed, genomes contain the set of genes, which can, under specific circumstances, be activated during the life of an individual. As such, genomes are essential in defining the *potential for gene expression*. However, the understanding of gene activation dynamics requires the consideration, not only of the genome and the environment, both external and internal to the individual, but also of the representation and modeling of the intricate gene regulations coded into the molecular biology of the individuals. In this sense, *gene expression networks* become remarkably similar to neuronal networks, which are also characterized by facilitation and inhibition between the variables (the "genes"). At the same time, a better understanding of gene activation and animal development is paving the way towards a more comprehensive understanding of the largest and most sophisticated neuronal network: the human brain. Indeed, a substantial portion of current research in neuroscience targets or at least involves the associated molecular and genetic aspects. Needless to say, many such investigations now involve the extensive use of informatics, hence the name *neuroinformatics*.

This interplay between the advances in bioinformatics and neuroinformatics defines an interesting positive feedback between these two areas, which should, with time, lead to an exciting synergy between bioinformatics and neuroinformatics, which we shall call *neurobioinformatics* (see Costa, 2003b). As it turns out, such integration is most welcome, as it is badly needed in order to enhance the nearly saturated human intellectual abilities. While in the short

term our intelligence is likely to be augmented by informatics - e.g., access to databases and artificial intelligence software through cameras and projectors installed in glasses, or even through direct bionic implants in the nervous system, advances in the long term may include the genetic redesign and improvement of the architecture of our central nervous system. There will be no more excuses for forgetting the birthday of your neighbor.

Human resources formation

The first important point to note about human resources is that this is, by far, the most important element in science. Indeed, while every piece of equipment is ultimately destined to become obsolete, well-trained and intended human beings are destined to, like wine, become better and better with time. At the same time, a well-formed scientist or technician acts as a source of training herself, therefore establishing a multiplicative system of knowledge transmission.

However, given the myriad of areas involved in bioinformatics, a second important issue arises, namely: *how to train bioinformatics professionals?* Should a mathematician become a biologist or vice-versa? While we wait for the fulfillment of neurobioinformatics, which may ultimately give us the required almost unlimited intellectual ability, the most effective way to conduct multidisciplinary research is through integrated teams of scientists and technicians who are trained in complementary areas but who also share a basic language allowing them to communicate the problems and work together on the interpretation of the results. Still, the question of how to acquire such a common knowledge remains. Of particular importance is how to provide such knowledge to people (grads or undergrads) with predominant basic training in the biological and exact sciences? Of course, we need to teach some math to biologists and some biology to exact scientists. But to what extent should we do this? While there is no doubt that these two branches of science are equally difficult (or easy), exact sciences have a more *vertical* nature, in the sense that the learning of one specific subject (e.g., complex variable calculus) requires a previous acquaintance with a long chain of preliminary subjects (e.g., linear algebra, real calculus and complex variables). Biological sciences, in turn, are characterized by a more *horizontal* nature, implying a wide range of knowledge (e.g., organic chemistry). Therefore, it appears to be the case that familiarization of biologists with exact sciences should start as soon as possible along the training program. On the other hand, it is interesting to expose exact scientists to a broad perspective of the biological world. After all, the application of informatics in biology is likely to quickly extend to all biological areas. In this sense, the perspectives for employment as a researcher or practitioner will always be enhanced, not only by an in-depth knowledge of some areas of expertise, but also by the flexibility and generality of knowledge in complementary areas.

Research and industrial perspectives

The vast prospects for profits stemming from bioinformatics have implied from its beginnings that a good portion of the research aimed at short term impact (i.e., issues currently leading to strong commercial implications with great potential for short term investment return) would be performed by research institutes associated with and/or maintained by major companies, to the point that it becomes a difficult challenge to compete (or complement) research done

by such consortia. A direct consequence of such a state of affairs is that the sponsoring companies will register many of the findings as patents. So, while academia and governmental research institutions should keep involved with such developments, their chances of success are more likely in longer-term research projects. An interesting alternative approach for the smaller-sized research institutes and companies is to try to identify issues that are currently of little interest but which may, in the medium term, transform themselves into important trends. In other words, great creativity is required from everybody. The public availability of data and results, accessible via internet, should continue to play an essential role, not only for catalyzing discoveries, but also for facilitating research for both small research teams and developing countries.

Ethics and moral issues

Although corresponding to the most important issues in science and humanity, relatively little attention has been focused on ethics and moral issues. Currently, one of the major approaches to ethics and moral states that scientists should concentrate on the development of scientific results, leaving its use for others (typically politicians) to decide. However, would not those issues, subjective as they may be, constitute themselves a primary object of interest of science? We should not forget that, after all, science is *from* humans *to* humans, or to what humans may in the end become.

CONCLUDING REMARKS

Early on, the human need for future prediction and the understanding of nature was to a great extent dependent on oracles, among which Delphi was the most famous. Being highly ambiguous, oracles could almost always be interpreted as being correct, a strategy that is still widely adopted in politics. The advent of science, based on the use of mathematics, paved the way to more objective and concrete predictions (models). Along these lines, I have discussed a few aspects related to the future of the new science of bioinformatics. As I did not use the same artifice as did the oracles, most of the trends that I have identified are very likely to be proven wrong.

It has been said that before the construction of the temple of Apollo at Delphi, people used to go to that very same place, high on the Parnasum mount, in order to hear messages about the future blown by the frequent wind coming from the sea. There, in an outstanding natural setting, humans spent hours listening, not really to the wind, but to their own nature and anxieties projected and amplified into the whistle of the wind. The future had already been predicted and had started to be accomplished: to know thyself. Modern science, through bioinformatics, is poised not only to substantially contribute to that cause, but also to reshape humans in the process. It has never been so important to hear the wind and reflect about what we want and expect. Only so will we be able to play a sensible active rule in the whole process.

REFERENCES

Albert, R. and **Barabási, A.L.** (2001). Statistical Mechanics of Complex Networks. *Rev. Mod. Phys. 74*: 47-97.

Bose, I. (2002). Biological networks (http://arxiv.org/abs/cond-mat/0202192).

Costa, L. da F. Costa (2003). A maçã e o biscoito da sorte. *Rev. Bras. Ensino Fís. 25*: 164-168. Available at http://www.scielo.br/pdf/rbef/v25n2/a05v25n2.pdf (in Portuguese).
Costa, L. da F. (2003a). The hierarchical backbone of complex networks. *Phys. Rev. Lett. 93*: 098702, 2004 (cond-mat/0312646).
Costa, L. da F. (2003b). Neurobioinformática: Quanto mentes e genes se encontram. Comciência, UNICAMP, http://www.comciencia.br/reportagens/bioinformatica/bio11.shtml (in Portuguese).
Costa, L. da F. (2004). A generalized approach to complex networks (cond-mat/0408076). Available at http://xxx.lanl.gov/abs/cond-mat/0408076.
Costa, L. da F. and **Cesar, R.M.** (2001). *Shape Analysis and Classification: Theory and Practice*. CRC Press, Boca Raton, FL, USA.
Duda, R.O., Duda, P.E., Hart, P.E. and **Stork, D.G.** (2000). *Pattern Classification*. Wiley Interscience, New York.
Erdös, P. and **Rényi, A.** (1959). On the evolution of random graphs. *Publ. Math. 6*: 290-297.
Holme, P.M., Huss, N. and **Jeong, H.** (2003). Subnetwork hierarchies of biochemical pathways. *Bioinformatics 19*: 532-538.
Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. and **Barabasi, A.-L.** (2000). The large-scale organization of metabolic networks. *Nature 407*: 651 (cond-mat/0010278).
Jeong, H., Mason, S.P., Barabási, A.-L. and **Oltvai, Z.N.** (2001). Lethality and centrality in protein networks. *Nature 411*: 41 (cond-mat/0105306).
Newman, M.E.J. (2003). The structure and function of complex networks. *SIAM Rev. 45*: 167-256 (cond-mat/0303516).
Vazquez, A., Flammini, A., Maritan, A. and **Vespignani, A.** (2003). Global protein function prediction in protein-protein interaction networks. *Nat. Biotech. 21*: 697-700 (cond-mat/0306611).
Wagner, A. and **Fell, D.A.** (2001). The small world inside large metabolic networks. *Proc. R. Soc. Lond. B 268*: 1803, 2001. Also: Working Papers of the Santa Fe Institute, 00-007-41.
Watts, D.J. and **Strogatz, S.H.** (1998). Collective dynamics of small-world networks. *Nature 393*: 440-442.