BIOINFORMATICS

Vol. 16 no. 3 2000
Pages 187–189

Editorial

TWO OR THREE MYTHS ABOUT BIOINFORMATICS

I could have started by describing the recent explosion in the analysis of biological information using computational approaches, how important the field of bioinformatics has become over the past few years and the like, but I will not. The chances are that even if you are reading this journal for the first time, you must have heard something about the field and therefore I consider this introduction unnecessary. What I would like to discuss here is a set of less obvious truths that unfortunately do not appear to be clearly understood by the majority.

This new field that is now known as *bioinformatics* has excited academic boards and private investors alike. Bioinformatics has provided the basis for the future of large-scale biology: it is relatively inexpensive, research and development can be done with modest equipment and plenty of public resources, you can produce and publish results rapidly, it can generate predictions that should be tested in the lab, and last, but not least, this is a novel and exciting technology platform for the future. I will now take the opportunity to discuss the above (mis-)conceptions and examine how much truth they actually contain. These innocent-looking myths create problems in the field that can have fateful consequences, and they warrant some more attention by the community.

MYTH #1 : "Anybody can do this!"

COROLLARY 1. *Bioinformatics is inexpensive.* With so many plentiful resources, any person equipped with an introductory textbook in biology, the web and a PC online, is in principle a bioinformatician in the making. In a more realistic setting, a person with an overview of key resources and in-depth knowledge of ten `unix` commands in any biological laboratory becomes the *de facto* bioinformatics support person for the whole group. Even in professional settings, this concept of providing for people equipped with a workstation dominates the funding model. This deceptive readiness of entering the field permeates the whole structure of bioinformatics, including various industrial set-ups and research funding agencies. Costs associated with professional software development environments, adequately planned and properly structured disk space resources, number and speed of CPUs available to any given person, projected costs of activities over time (usually growing exponentially to follow the data), and in general the proper infrastructure for a professional practice of the field, are not always fully appreciated. As a matter of fact, the relatively inexpensive nature of bioinformatics is only applicable in pharmaceutical research, where it can cut corners in the long run towards drug target selection and identification, previously done in a more *ad hoc* (and expensive) manner. Probably this myth has its origins in academia, which conveniently borrowed the concept from industry. The truth is that bioinformatics is an expensive activity, being merely cost-effective if compared to previous practices in pharmaceutical research. This is a serious issue that needs to be addressed: bioinformatics neither strictly follows the financial requirements of biological research nor is it just a biological field 'without laboratory consumables' (a phrase I have heard often). Given the recent sharp twists and turns in the genomics arena, it is important to stress that serious long-term commitment of resources is required to bring any bioinformatics project(s) into fruition.

COROLLARY 2. *Bioinformatics software is free.* I have observed that many people consider that they can 'do' bioinformatics as a side project. This is related to the above remarks. The ease of accessibility of resources, the multitude of web services for biologists, the number of public-domain databases, and the low cost of computers, suggest that essentially any scientist or engineer can become a bioinformatics practitioner. This may be partially true, and indeed some of the work can be done with such ease these days, that physical location is increasingly irrelevant, compared to other biological disciplines. However, this potential should not be over-emphasized and the problems that are created by this myth should be sufficiently addressed. One of the most serious issues is that the effort (time and money) for software design and development is not sufficiently appreciated. Sometimes, there are dependencies on commercial software that you need to buy (yes, buy) to develop layers of applications software. The general issues of the software life-cycle, including design, development, testing, maintenance, quality control and documentation are not usually part of a cost equation. The rate-limiting step may indeed be the quality of software. Surely you can write a `perl` script that does something interesting, but that does not mean you have suddenly became a bioinformatics expert. Many software systems are out there, and their shortcomings are always attributed to insufficient resources. Instead people should be asking whether the software design and testing was adequate. One wonders what percentage of currently available software systems have been developed by DIY bioinformatics experts. It must be appreciated that not only quantity but also quality of software is an essential and expensive element of this field.

MYTH #2 : "You'll always need an experiment!"

COROLLARY 1. *Bioinformatics is a rapid-publication field.* Experimental biologists envy the speed and effectiveness of computational biology publishing. They may be thinking, "all these people do is push a button and spend the rest of their time writing papers". Nothing could be further from the truth. First, the set-up of any computation is a very expensive operation and has to be done with great attention to detail and procedures, very much like an experiment. Second, there is a lot of background work that is going on, and which is (or cannot be) published in the traditional biological literature. For instance, if you have been developing a database system during the recent years, your best bet would be a computer-science proceedings publication. To me, the field is just more productive in real terms (or in terms of publications) because it speeds up observation and performs analyses that are impossible to perform using experimental approaches. In a sense, it represents a case where new technology supersedes part of the tradition. The more biologists become computer- literate, the more the riches of their own experiments will be revealed for all to enjoy, and bioinformatics specialists will devote most of their time to develop better systems to support experimental research. In the meantime, and given the current landscape, computation (ideally coupled with experiment) will still provide the fastest and most comprehensive way of performing research. Scientific papers, though, will still be written in their usual, time-consuming and laborious ways.

COROLLARY 2. *All bioinformatics does is generate testable predictions.* This is possibly one of the most controversial points in this essay. There is a wide belief that computation in biology 'supports' experimental research, where all truth lies. This is an epistemological issue, and I don't have enough space to elaborate. However, taken at face value, bioinformatics for biologists is a set of tools that help you reach the ultimate truth. I could not disagree more. The argument goes like this: computation makes predictions, some of which sound interesting and, depending on other information (or the lack thereof), can be tested in a 'wet' experiment. In molecular biology, there is usually a hunch, which is implicitly formulated as a hypothesis, then there is an experiment that 'proves' (or rather, supports) the initial guess. In computational biology, something similar happens: there is a null hypothesis (e.g. a certain sequence has no homologues in the database), then there is the experiment (e.g. search the database) and then there is the test, which rejects or accepts the null hypothesis (e.g. sequence does or does not have homologues). This is a precisely formulated process, and hence a good experiment. In other sciences, computation is regarded as a significant contribution towards a deeper understanding of the subject matter, but not in biology. On the other hand, experimental work can be error-prone to considerable degrees. I could provide anecdotes where a closer look, a lab Macintosh and the internet, would indicate that the conclusions from a given experiment were incorrect—but there is no space. I may be exaggerating here to make the point, but more attention should be paid to the inherent nature of computation as experiment, and its scientific rigour in supporting, or even contradicting, experimental information.

And finally comes the third myth, which in my opinion is the most serious of all:

MYTH #3 : "This is a new technology, but technology nevertheless!"

COROLLARY 1. *Bioinformatics is a new field.* This is one of the first and foremost misconceptions: people new to the field think the field is new. Computation in biology has been a dream for almost a century. Back in 1924, Lotka said, in his 'Elements of Physical Biology': "It remains [] to enumerate the methods by which Physical Biology may be expected to develop. For the *gathering* of data two methods are available: observation in natural conditions, and observation under experimental (laboratory) conditions. [. . .]. For the *elaboration* of data, the establishment of regularities (laws), there is available in this field, as everywhere in science, the method of *induction*, aided, if need be, by statistical technique [. . .]" (*verbatim*, incl. italics). Since then, scientists tried to quantify and measure properties of biological systems, with little or no success, because biological matter refused to be quantified. Only when measurable biological quantities started appearing in the form of sets of three-dimensional co-ordinates, or later, as strings of characters, then computation in biology boomed. Early successes in computational *molecular* biology include the works of Volkenstein, Pauling, Dayhoff and other pioneers. When in the 1970's, the first algorithms and, most importantly, their computer implementations became available, biological computation and theoretical research in biology started being shaped towards its present state. The rest, as they say, is history.

COROLLARY 2. *Bioinformatics is an applications discipline.* This is probably a fatal flaw. It is highly unfortunate that this term was coined during the parallel explosion of the internet for the masses. It contains a highly suggestive connotation that bioinformatics is purely an applications discipline, a field that provides solutions for biological sciences, with mere application of existing technologies borrowed from 'informatics', a term much less frequently used in computational science. The term has plagued in some ways the field of bioinformatics, by strongly suggesting that this is a technological endeavor with clear solutions (e.g. algorithms) in need of sheer development. I

do not subscribe to this view. There is serious intellectual content in this field, with a range of challenging scientific questions that remain to be answered. Despite the (almost elusive) progress, all major problems in the field remain unsolved. Some examples include biological questions (such as how does molecular function evolve) and computational questions (such as what is the best way of inter-operating database systems). By approaching bioinformatics (or computational biology) as a science, the field will not be misapprehended as a technology platform, where people know the problems whose solutions are only a matter of time and resources. Instead, it should be viewed as a genuine scientific discipline where vivid imagination and fleeting serendipity take equal stance along careful planning and precision engineering.

Conclusions

My motives here are well-intentioned: only when this field is viewed as a proper scientific discipline, with its intellectual depth and substantial history, and on a par with experimental biology, some of the above myths will be dispersed with positive effects for the scientific community. Bioinformatics currently represents a wide range of activities, there is already a vast amount of literature associated with it and a clear and ever-increasing need for highly specialized staff. These are promising signs but they have to be properly supported by adequate changes in the biology curriculum, both at the undergraduate and graduate levels, with more computation and quantitative approaches, radical changes in the funding system to address the specialized needs of the field in terms of long-term support for staff in a highly dynamic job market, and projected costs for resources, including software systems, and finally an increased awareness of the value that computational approaches can represent for the biology of this millennium.

My thanks to Barbara Cox, Antoine Danchin, Peter Karp, Liisa Holm, Chris Sander and Alfonso Valencia, for valuable comments.

Christos Ouzounis
Associate Editor,
Group Leader, Computational Genomics Group Research Programme
The European Bioinformatics Institute
EMBL Cambridge Outstation
Wellcome Trust Genome Campus
Cambridge UK