

Editorial

BIOINFORMATICS AND THE THEORETICAL FOUNDATIONS OF MOLECULAR BIOLOGY

To continue on some issues raised previously (Ouzounis, 2000), I would like to examine the principal elements of bioinformatics and its relationship to biological sciences. In various social settings, over conference dinners, at annual retreats and other occasions, I have been involved in these types of casual debates. Is bioinformatics a science or a mere technology platform? Or, put more bluntly, where is the science in bioinformatics? If it is a science at all, what are its foundations and key steps in its development? What is the position of computation in an experimental science? These are the epistemological issues I would like to explore further in this editorial.

In some ways, the term 'bioinformatics' is a very successful one. It reflects the theory and practice of computing behind all aspects of biological sciences, from 'pure' algorithm research to 'applied' support for experiment. At any rate, bioinformatics has been a tremendously successful discipline, thanks to a powerful combination of an explosive growth in the computer technologies at large with the streamlining of biological experimentation, in what has been called high-throughput biology (genomics, proteomics and the like). Although this success is not directly measurable, there is no doubt that modern bioinformatics has contributed to the development of various sub-disciplines in biology and the increase of our knowledge of biological systems, their structure, function and evolution.

However, the usage of the very same term can be slightly more controversial when applied to more fundamental aspects of biological science. For example, comparing hundreds of genome sequences in order to identify conserved gene structures may indeed pose a significant challenge to current computational methods but it may not be considered by some as part of bioinformatics research. It is "a biological problem", merely addressed by computer methods. This is where the problems begin. It is ever so common to identify all technical aspects of computational research in molecular biology (i.e. research without 'wet', "real" experiments) as 'bioinformatics', because of the strong connotations of this term. At the same time, this 'tech' designation sometimes denies the discipline access to its own fundamental scientific issues. This may be the reason why a number of workers have extensively used the term 'computational biology', possibly with the aim to distance themselves from and part with the term 'bioinformatics'.

So, we may end up with a situation where all the

'good stuff' of fundamental biological research are called anything-but 'bioinformatics', further implying that the infrastructure support and technical work is not scientifically very exciting. This is a common scenario, I believe, in many quarters. However, things are not that simple. Development of key methods has always been inextricably intertwined with the analysis of complex data and the deeper understanding of a key biological question. Fundamental aspects of biological science such as molecular sequence homology, protein structure engineering and design, species phylogeny and taxonomy, and the analysis and simulation of molecular networks have all been addressed and explored by bioinformatics research.

Seeking a loose analogy with other sciences, one may suggest that purely technical aspects of any science may be given a corresponding name and are considered as a field on their own, for all practical purposes. For instance, a key technology platform for quantum physics is 'physical optoelectronics'. Nobody would ever suggest that a technological breakthrough in physical optoelectronics would necessarily be considered as an advance in quantum physics, and vice versa—unless theory and application are involved in an intimate interplay! Thus, I believe that there is much more to bioinformatics than meets the eye. This is a fertile field, which incubates the birth of a new kind of biology, the data-driven, inductive analysis and simulation of biological matter. I maintain that the field of bioinformatics (despite the somewhat unfortunate naming!) forms the basis for a truly theoretical biology, upon which our knowledge and ever increasing understanding of biological systems at the molecular level is being built (Ashburner *et al.*, 2000).

It is too daunting a task to describe the history of the field from its humble origins to its current position in the midst of the genomics revolution. Pioneers of the field have given their personal accounts of the early days much more eloquently than I could possibly have done (Sander *et al.*, 2000). The point here is that the history of the field goes all the way back to the very beginnings of molecular biology and that computation has been lurking around much longer than people usually appreciate. It is hard to give precise milestone dates for key developments, but one could mention the advent of sequence alignment algorithms (early 1980s), community-driven availability of data and databases (mid 1980s), rapid database search systems (late 1980s), sophisticated protein structure prediction systems (early 1990s), genome annotation and comparison (mid 1990s) and functional genomics analysis systems (late 1990s).

The answer to the original question of whether bioinformatics is indeed much more than a technology platform (and possibly an important scientific field!) should

be strongly affirmative, in my opinion. Computation in biology is a key element in modern research. At the same time, experimental biology and its subject areas use bioinformatics along with other technology platforms (e.g. genetic engineering or electron microscopy) to address important scientific questions.

Once we have adopted the above arguments, we can move on to the more challenging aspects of the problem, to examine how computational biology fares in comparison with the supremacy of experimental biology. In some ways, the whole question amounts to the relationship of computation to experiment in a predominantly experimental science. Computational scientists consider their sophisticated and meticulous calculations as genuine experiments, with proper controls, and hypotheses to be tested. However, there is a widely held view that experimental scientists usually look down at computation as a supplement of (but rarely a replacement for) experimentation. To what extent is this true? And how can computation win the hearts and minds of experimental scientists?

The question can be rephrased as follows: what is the role of computational work or theoretical research in a primarily experimental science, such as physics or biology? I can think of two main reasons: first, the induction of general laws from the synthesis of various experimental observations; and second, the application of these laws to most situations without the use of experiment - unless these laws prove insufficient. What are the general laws that bioinformatics and computational biology have produced for biological science? Theoretical research in molecular biology using computers has indicated that:

- three-dimensional shapes of protein molecules are more conserved than their biochemical functions;
- most point mutations in protein molecules occur on the molecular surface;
- genome sequences are highly related and truly reflect species phylogeny; gene numbers, however, do not.

These clauses may be considered as general laws for biological systems because there is a multitude of corollaries deriving from them with direct applications for biological research. For example, identifying a homologous sequence 99% identical to a protein of known function should imply that the differing residues should normally

have little effect on structure, would be present on the surface of the molecule and have no significant effect on biochemical function; thus, eliminating experiments addressing any of the above issues. Thus, 'laws' of this kind can accelerate biological research by (i) organizing knowledge and (ii) replacing experiment (where possible).

Those who still doubt the validity of computational biology as a genuine, stand-alone scientific field, usually resort to a comparison with computational physics and point out the lack of quantitative models of wide generality in biology. It is true that biological computation has not reached the high degree of sophistication that theoretical physics has achieved. Yet, the nature of computing with biological matter is radically different from that of the physical world, supported with complex databases, qualitative reasoning and symbolic computation (Karp, 2001). In some ways, many experiments are being made redundant or irrelevant, thanks to the powerful mix of prior knowledge plus the inescapable data analysis under our general 'laws'. The predictive capability of current bioinformatics systems is becoming formidable, eliminating the need for more mundane experiments. If still in doubt, next time you think up of an experiment, consider how many experiments you have already eliminated thanks to computation in biological research. And this, not only thanks to user-friendly software or fast computers, but also thanks to our acquired knowledge accumulated over all those years during which we have been computing with biological matter.

REFERENCES

- Ashburner, M. *et al.* (1995) Gene ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.
- Karp, P.D. (1995) Pathway databases: a case study in computational symbolic theories. *Science*, **293**, 2040–2044.
- Ouzounis, C. (2000) Two or three myths about bioinformatics. *Bioinformatics*, **16**, 187–189.
- Sander, C. *et al.* (eds) (2000) History issue. *Bioinformatics*, **16**, 1.

Christos Ouzounis
Computational Genomics Group
The European Bioinformatics Institute
EMBL Cambridge Outstation
Cambridge CB10 1SD UK
Email: ouzounis@ebi.ac.uk
www.genomes.org