Emerging Topics in Life Sciences (2017) **1** 241–243
https://doi.org/10.1042/ETLS20170137

Editorial

Integrating the whole from the sum of the parts: vignettes in computational biology

Jeffrey Skolnick

Center for the Study of Systems Biology, School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332, U.S.A.

Correspondence: Jeffrey Skolnick (skolnick@gatech.edu)

As is typical of contemporary cutting-edge interdisciplinary fields, computational biology touches and impacts many disciplines ranging from fundamental studies in the areas of genomics, proteomics transcriptomics, lipidomics to practical applications such as personalized medicine, drug discovery, and synthetic biology. This editorial examines the multifaceted role computational biology plays. Using the tools of deep learning, it can make powerful predictions of many biological variables, which may not provide a deep understanding of what factors contribute to the phenomena. Alternatively, it can provide the how and the why of biological processes. Most importantly, it can help guide and interpret what experiments and biological systems to study.

As is typical of contemporary cutting-edge interdisciplinary fields, computational biology touches and impacts many disciplines ranging from fundamental studies in the areas of genomics, proteomics transcriptomics, and lipidomics to practical applications such as personalized medicine, drug discovery, and synthetic biology. The realization of successful predictive computational biology algorithms employs many state-of-the-art techniques, such as dynamic programming, Brownian and molecular dynamics simulations, and more recently deep learning, and is fundamentally multiscale in nature. Since biological systems contain millions of interacting molecules and assemblies on distance scales ranging from Angstroms to meters and time scales ranging from picoseconds to years, it is a prototypical big data field. This issue in *Emerging Topics in Life Sciences* contains three articles that deal with cutting-edge facets of this field that personify its diversity. To understand the behavior of a given organism, one at a minimum needs the parts list. The article 'Analysis of single-cell genome sequences of bacteria and archaea' by Bowers, Doud, and Woyke [1] details recent progress in the sequencing of the DNA of individual bacterium or archaeon. This then provides the parts list, making up a given bacterial or archaeal genome. This is important as many bacteria cannot be cultured in the laboratory; as a consequence, their genomes are unknown. One could, of course, perform metagenomics and collect the DNA sequences of the soup of many species, but far greater information is provided when one knows the individual genomes. The same idea holds for cancer tumors which also contain a heterogeneous population of diverse cancer cells. Just as mixing blue and yellow paint colors makes green, looking at a jar of green paint one would not know if it is really a mixture of blue and yellow paint molecules, none of which are green or if all the molecules are green. Similarly, looking at the composite DNA of many species does not provide the necessary segregation to enable the understanding of the biochemistry and physiology of the individual bacterial species and how they are related to each other. Of course, even knowing the DNA sequence of an organism is but the first step toward understanding its biology. Indeed, the complexity of an organism is not dictated by its DNA sequence alone. As pointed out in the Commentary by Ponting titled 'Big knowledge from big data in functional genomics' [2], the ultimate goal of science is to exploit genomics data to provide deep insight beyond a parts list. Among the many questions one wishes to address are the following: What genes code for which proteins and what do the proteins do, how do all the millions of molecules within a cell interact, and what are the consequences of mutations in the proteins and RNA on the health and

Received: 28 September 2017
Revised: 5 October 2017
Accepted: 5 October 2017

Version of Record published:
14 November 2017

Emerging Topics in Life Sciences (2017) 1 241–243
https://doi.org/10.1042/ETLS20170137

physiology of the organism? Ideally, one would like to be able to understand how a given cell works and if it is malfunctioning, and how it could be fixed. Thus, biological insights can operate on many levels. Given the plethora of biological data, one can use the very powerful tools of deep learning to infer associations and make predictions. The contribution by Jones, Alasoo, Fishman, and Parts titled 'Computational biology-deep learning' gives an excellent overview of what deep learning is and some of the many fields it has been applied to with remarkable success [3]. It also provides a very nice roadmap of what various deep learning computational tools are designed to do. The advantage of deep learning is that it readily integrates a massive quantity of diverse types of big data and can make quite complex predictions when appropriately trained. For example, it can be used in image interpretation, the prediction of protein subcellular location, the prediction of gene expression, DNA methylation and replication timing, or relevant for this editorial, the prediction of RNA secondary and tertiary structure. However, for all its tremendous power, it is very much a black box, and rarely does it provide deep insight into how or why. One example of how deeper understanding could be obtained is provided in the article by Thiel, Flamm, and Hofaker, 'RNA structure prediction: from 2D to 3D' [4]. DNA codes for RNA which not only codes for proteins using the transcription machinery of a cell but also plays important regulatory roles. To understand, at least partly, how this occurs and how RNA actually works, it is very helpful to have the three-dimensional structure of the RNA molecule itself. The spatial arrangement of the RNA base pairs can suggest how a given RNA interacts with other small or large molecules in the cell. Note that the field of RNA secondary and tertiary structure prediction is less mature than its protein counterpart. This is at least partly because there are far more experimentally determined structures of proteins than there are of RNA molecules. To develop a robust prediction algorithm requires the existence of large representative data sets so that training and testing can be segregated. This is necessary in structure prediction in particular and deep learning in general to ensure that the given algorithm generalizes rather than simply memorizes the input data. While the situation for RNA structure prediction is improving, at present, it is a more nascent than the protein structure prediction field. It is quite likely that the tools developed to predict protein structure, when appropriately translated from amino acids to nucleic acids, could help advance RNA structure prediction. In that regard, methods of RNA structure comparison that assess whether a given structure prediction is statistically meaningful or random need to be introduced into the field of RNA structure prediction. Yet, even if fully successful, knowing the structure and function of the individual molecules does not necessarily give insight in the collective behavior of cells. The macromolecular environment within a cell is much like a crowded party in Times Square on New Year's Eve rather than an isolated, non-interacting set of molecules. It is this collective behavior that gives rise to living systems. What is apparent is that computational biology can play a major role in advancing biological understanding. By integrating massive quantities of biological data, it can identify collective features and key biological processes that could then be experimentally validated. Computational biology plays a role in the generation of the genomic data itself by helping to identify the genes and their variations. Indeed, it provides the tools for genome assembly and plays a key role in functional annotation. It can suggest what are the most relevant experiments that should be done to elucidate key biological principles. That is, it can suggest to an experimentalist, 'Look here, not there'. Thus, it is an invaluable partner to most, if not all, aspects of experimental biology as it transforms from a descriptive to a quantitative science.

Summary

- Computational biology plays a key role in all aspects of modern biology.
- Its utility is widespread and ranges from helping assembling parts lists of cells, integrating big data by powerful machine learning tools and provides guidance on the how and why biological processes occur.
- Examples of areas where computational biology is important include the analysis of single-cell genomes, the prediction of RNA secondary and tertiary structure and the application of deep learning to a plethora of areas including image processing, the prediction of protein subcellular location, gene expression, DNA methylation, replication timing, and RNA secondary and tertiary structure.

Funding
This work was supported, in part, by [R35GM118039] of the Division of General Medical Sciences of the NIH.

Acknowledgements
I would like to thank Dr Anton Enright (EMBL-EBI) for his input into the initial discussions around the concepts covered in this issue of *Emerging Topics in Life Sciences*, and for securing some of the contributors and content.

Competing Interests
The Author declares that there are no competing interests associated with this manuscript.

References
1 Bowers, R.M., Doud, D.F.R. and Woyke, T. (2017) Analysis of single-cell genome sequences of bacteria and archaea. *Emerg. Top. Life Sci.* **1**, 249–255 https://doi.org/10.1042/ETLS20160028
2 Ponting, C.P. (2017) Big knowledge from big data in functional genomics. *Emerg. Top. Life Sci.* **1**, 245–248 https://doi.org/10.1042/ETLS20170129
3 Jones, W., Alasoo, K., Fishman, D. and Parts, L. (2017) Computational biology: deep learning. *Emerg. Top. Life Sci.* **1**, 257–274 https://doi.org/10.1042/ETLS20160025
4 Thiel, B.C., Flamm, C. and Hofacker, I.L. (2017) RNA structure prediction: from 2D to 3D. *Emerg. Top. Life Sci.* **1**, 275–285 https://doi.org/10.1042/ETLS20160027