Editorial

Michael P.H. Stumpf^{1,2}

Biology challenging statistics

¹ School of Biosciences and School of Mathematics and Statistics, University of Melbourne, Melbourne, Australia, E-mail: mstumpf@unimelb.edu.au

² Department of Life Sciences, Imperial College London, London, UK, E-mail: mstumpf@unimelb.edu.au

DOI: 10.1515/sagmb-2018-0048

Biological research has traditionally been a observational and experimental endeavour. Modern biology, with its reliance on genomics, proteomics, transcriptomics is phenomenally data-rich. In the near future biological data may exceed the vast amounts of data generated in other sciences, including astronomy and high-energy/elementary particle physics.

Over the past century or so, the biomedical and life-science communities have – sometimes grudgingly, sometimes enthusiastically – embraced quantitative, statistical and computational approaches to make sense of observations and data. *SAGMB* has always aimed to capture technical innovations in the realm of statistics applied to biological problems and data. This is a nontrivial endeavour: we are trying to serve the community of statistical and machine learning innovators and method developers, but also reflect, and ideally shape the tools and techniques employed by the genomics, biomedical, and systems biology communities. Technical innovation and sophistication does not always translate into useful or practical tools. And some widely used approaches in e.g. biomarker discovery or association studies are popular, because they are easy to use and understand, and not because they are technically superior to other approaches (and frequently they are not).

SAGMB has an important role to play in providing a world-class forum for methodological research, and the past few years have seen a shift in approaches away from e.g. population genetics to more high-throughput data analysis and systems biology. This is an area of continuing (and, in my view urgent) need of methodological rigour; rigour that should include definitive comparisons of different approaches, so as to guide community efforts into a direction of better replicability of data analysis, and better understanding of the limitations of e.g. machine learning approaches. The somewhat slower pace of the statistical literature, compared to molecular biology and genomics publications, does not have to be a disadvantage: hypes and fashions come and go, and it is fascinating to watch e.g. how “deep learning” is currently developing: at the moment it is anyone’s guess as to how widely applicable the veritable zoo of deep learning approaches will ultimately be, but already we are seeing ways in which we can understand the performance and power of these approaches in terms of tried and tested (and very well understood) more conventional statistical methodologies, such as regression (of course) and adaptive basis function approaches.

One particular area for which I have high hopes for *SAGMB* in the future is to bridge the gulf between mechanistic and data-driven modelling. A lot of potentially exiting work can be done at trying to combine such approaches. In the context of disease genetics, for example, we still rely to a large extent on black-box statistical models. There have, however, also been some exciting initial studies that suggest that mechanistic models (using e.g. information about the structure of gene regulation and protein-protein interaction networks) will allow us to dissect in more detail the molecular basis of disease phenotypes. Early results have suggested that disease traits are *omnigenic*, i.e. all genes contribute, albeit to different degrees. The molecular interaction networks mediate the effects of genes, and allow us to reason about relative contributions resulting from different gene products or pathways. This is a potentially exciting problem that combines many of the techniques that have been at the core of *SAGMB*’s mission.

From a data-rich science, biology has increasingly been turning into a hypothesis-rich science. Hypothesis testing and (perhaps preferable) model selection should thus be taking centre stage and a host of fascinating problems remain to be solved: correcting for multiple hypotheses in the presence of correlations, Bayesian model selection approaches for large model spaces, and prior elicitation for model selection problems are only some of the examples. Model checking is another field which I view as increasingly important: predictive power of a model alone is not a robust way to assess a model’s strengths or inadequacies, and finding ways of probing a model decisively is statistically challenging, and frequently requires considerable levels of domain expertise.

Some of the problems that are being investigated in the modern life-sciences are of enormous complexity: the development of a multi-cellular organism from a single fertilised egg cell is one such example. Being able to probe mRNA abundances at single cell resolution in an organisms is now becoming possible, and statistics

Michael P.H. Stumpf is the corresponding author.

©2018 Walter de Gruyter GmbH, Berlin/Boston.

and machine learning approaches in all their guises are frantically applied to such data. Being able to derive mechanistic insights into e.g. developmental programmes, from such data will remain a fruitful and exciting field for statistical research.

I am confident that our new *Editor in Chief*, Guido Sanguinetti, will provide the vision and leadership to keep *SAGMB* at the forefront of this exciting area of research. He has ample and wide reaching expertise in combining state-of-the-art statistical methodology and applying it to exciting problems in biomedical research. More importantly, he is also an extremely lucid writer and communicator. I wish him success and joy in this new role. And I am delighted that he has taken on this role.