Mathematical and statistical challenges in bridging model development, parameter identification and model selection in the biological sciences

R. E. Baker (University of Oxford),
D. Coombs (University of British Columbia)
M. J. Simpson (Queensland University of Technology)

November 11 – November 16, 2018

1 Overview of the field

Using mathematical models to assist in the design and interpretation of biological experiments is becoming increasingly important in biotechnology and biomedical engineering research; yet fundamental questions remain unresolved about how best to integrate experimental data within mathematical modelling frameworks to provide useful predictions. Traditional approaches incorporating mathematical and computational models in the design and interpretation of biological experiments often rely heavily on heuristic methods that vary from user to user, and from application to application. Not only is the selection of modelling frameworks often subjective, the integration of these modelling frameworks with experimental data is largely driven by individual preferences on a case-by-case basis. **Such variability in scientific practices is at best undesirable, and at worst leads to issues associated with research reproducibility**. Novel mathematical, statistical and computational tools are needed to provide a standardised pipeline that enables experimental data to be used effectively in the development of models, and in model parameterisation and selection.

One key challenge in using mathematical modelling to interpret biological experiments is the question of how to integrate multiplex, multi-scale quantitative data generated in experimental laboratories to improve our understanding of a specific biological question. These data might, for example, include time-series data for the concentrations of key intracellular biological signalling molecules, time-lapse microscopy data visualising the distribution of key cytoskeletal components, or tissue- and organism-scale data describing cell rearrangements or fluid flows. A standard protocol, that includes the design of experiments targeted towards parameterising models, validating specific model hypotheses and inference of underlying mechanisms based on quantitative data, is lacking.

To a large extent these issues are compounded by the fact that attempts to connect quantitative data with mechanistic models are being made in disparate fields of biology that function on a range of spatio-temporal scales: from the study of biochemical signalling networks within individual cells; to tissue-scale models of development, disease and repair; to ecological models of animal populations. At present, a major barrier to progress is a lack of cross fertilisation of ideas, or an awareness of the techniques and methodologies being developed in other fields of specialism. **Our observation is that the lack of cross fertilisation of ideas is because this area of research falls between many established disciplines, such as applied mathematics, applied statistics and computational mathematics, and this makes the development of consistent practices and protocols inherently difficult**. As such, the broad aim of this workshop was to bring together

researchers working in different areas of mathematics and statistics, and on different biological applications, to share and develop their research ideas towards bridging mechanistic model development, parameter identification and model selection using quantitative data.

2 Recent developments and open problems

Novel mathematical, statistical and computational tools are required to enable models to be developed, parameterised and validated as part of the predict-test-refine-predict cycle essential within biology. The overarching goal of the workshop was to work towards developing the mathematical, statistical and computational tools needed for a standardised pipeline that integrates multiplex, multi-scale quantitative biological data and mathematical and computational modelling. It is extremely timely because the recent explosion in the amount of quantitative biological and ecological data available means that the current disconnect between experimental sciences and mechanistic modelling will widen without interdisciplinary intervention. In addition, international meetings to discuss mathematical and statistical challenges in bridging model development, parameter identification and model selection in the biological sciences have been limited. The workshop focussed on tackling three main problem areas in the field.

Key challenge one: model coarse graining.

In order to resolve the conflict between requirements for including detailed descriptions of multi-scale biological processes into models and being able to efficiently simulate and/or subject models to analytical exploration, coarse graining can be performed. However, with biological data being generated on increasingly fine-grained scales, a significant challenge is to develop coarse graining approaches that retain descriptions of key processes happening across spatio-temporal scales [1, 13]. A relevant example is how to include detailed descriptions of tissue perfusion, that include descriptions of flow within blood vessels, into models that aim to understand tumour growth and, ultimately, determine optimal cancer treatment protocols.

Key challenge two: efficient methods for computational inference.

Computational inference methodologies target the posterior distribution (probability of the parameters given the observed data for a specific model), and are required when estimating models of biological systems as the likelihood function is generally analytically intractable (see *e.g.* [20]). There are significant challenges associated with computational inference that were discussed within the workshop. A key challenge, that was the focus of many discussions, is the development of understanding of when and where it is appropriate to use different types of inference methodologies, and how to quantify the errors associated with them.

Key challenge three: experimental design, model selection and uncertainty quantification.

Experimental design and data collection need to be optimised with respect to the specific biological or ecological question of interest and the model used to interrogate it [12]. The field currently lacks standardised approaches to model-driven experimental design. In addition, existing methods for model selection (encoding different biological hypotheses) can return different outcomes [18], and become difficult to interpret in the context of complicated models and noisy data, which is fast becoming the norm in the field. Methods for quantifying uncertainty include those that aim to understand how system outputs are affects by uncertainties in inputs such as parameter values (forward uncertainty propagation), and those that aim to measure discrepancies between data collected from a given experiment and the predictions of a mathematical model of that experiment (inverse uncertainty quantification). Uncertainty quantification is not routinely performed, however, the development and standardisation of methods to report and discuss uncertainty is key to the integration of modelling within the biological toolbox.

3 Presentation Highlights

The participants of the workshop included representation from Australia, Canada, New Zealand, Norway, United Kingdom, and the United States, with a range of career stages represented, from early-stage doctoral students, through postdoctoral research assistants, research fellows, tenure-track faculty and principle investigators. The participants were united by the use of common methodologies and ideas, that bring with them the same key challenges, on very different biological applications. As such, none of the participants, knew more than a handful of others in advance of the meeting. This common bond, and the small-scale nature of the workshop ensured that many new links and collaborations were formed; the meeting has really stimulated the development of a new research community.

Darren Wilkinson kicked off the meeting with an excellent talk outlining methods for Markov process parameter inference [21]. Darren compared some of the different "likelihood free" algorithms that have been proposed, including sequential ABC and particle marginal Metropolis Hastings, paying particular attention to how well they scale with model complexity.

Mattias Chung continued this theme, discussing the challenges in parameter inference for biological systems with noisy data, model uncertainties, and unknown mechanisms. Here, parameter and uncertainty estimation problems are typically ill-posed, meaning solutions do not exist, are not unique, or do not depend continuously on the data. Furthermore, experimentalists face a dilemma between accuracy and costs of an experiment. Mattias discussed new developments for parameter and uncertainty estimation for dynamical systems [8], as well as novel techniques for optimal experimental design [7], using three example applications.

Adelle Coster gave a very focused presentation about building models of glucose tolerance [6], with an emphasis on constructing useful mathematical models that capture biological processes at the level of a cell. In particular Adelle wishes to create models and compare predictions to experimental data so that she can rank the importance of various features and processes encoded in the data, and to make informed decisions about which model provides the "best" description of observations, by combining mean-field experimental data as well as observations of experimental noise.

Gary Mirams presented recent work regarding mathematical models of ion channel modelling with application to cardiac modelling [10]. Gary explained that to make progress in the field he needs to propose models, undertake an identifiability assessment to be sure that the models are identifiable given the types of data available and this leads to questions of optimal experimental design and questions of parameter inference, all of which feedback into the original question of model proposition. Gary's talk was used as the basis for a longer group discussion on the following day (see below).

Adam MacLean spoke about models of kidney development and inference, with a focus on branching processes during morphogenesis [11]. A key tool to connect models and experiments is approximate Bayesian computation, with a focus on identifying the key parameters in relatively complicated experimental data. Adam's model takes the form of an individual based model simulating stochastic cell migration and cell proliferation, with a continuous field of growth factor, which influences cell proliferation into neighbouring lattice sites giving rise to branching structures. Experimental data are movies from explants, giving rise to questions about comparing images from movies with images from simulations. Summary statistics are the area of the epithelia and the number of branches and ABC rejection is used to sample the posterior distribution. The main novelty is to use ABC rejection to generate an intermediate result and then refine the result using AABC, finding that branching is most sensitive to branching parameters.

Alex Browning spoke about identifying parameters in a continuum model of malignant invasion where malignant melanoma cells migrate, proliferate and degrade surrounding skin tissues [4]. Alex presented a typical continuum PDE model and a set of experimental data showing the invasion of the malignant population into the surrounding skin. Measurements of experimental noise allow the construction of an exact likelihood and Alex showed that invasion depth versus time is insufficient to distinguish between the three parameters in the model. Alternatively, Alex showed another two sets of experiments without skin tissues that

allowed him to simplify the model and use Bayesian learning to learn parameters from one experiment and apply them to the next experiment. Overall, the new approach leads to well-formed posterior distributions that agree with some previous parameter estimations.

Michael Plank presented a stochastic birth-death model developed to describe populations of plants with no movement, and the key effect is to incorporate spatially-driven competition so that individuals undergo a constant birth rate, but a density dependent death rate to reflect the impact of competition for nutrients and space. Stochastic simulations reveal how the spatial competition affects the model outcomes and standard mean-field arguments do not account for such spatial effects. The construction of a spatial moment model, whereby interactions between pairs of individuals are accounted for, was presented and the numerical solution of the governing equations provides a good match to the stochastic simulations [2]. A key aspect of the presentation was to explain how the ecology-based model with no movement might be extended to apply to a model of cells in which cells are able to undergo proliferation, death and movement events.

Rob Deardon spoke about building discrete time S-I-R models of disease spreading with the key aim of working in real time. Working in a Bayesian framework with an explicit likelihood, the main aim is that key features are not observed since real cases give no idea of the infectious period. Instead there are measurements of reporting time or notification time. This means the simulation times are an issue since the MCMC algorithm needs to cover the joint posterior of the model parameters, incubation periods, delay periods and parameters describing the delay periods. The main idea is to speed-up using a simpler simulation framework called an emulator, which Rob explained and showed promising results leading to significantly faster computation times [14].

Dennis Prangle gave a presentation about inference on stochastic differential equations using methods from machine learning, called variational inference, with the aim of obtaining results faster than MCMC [17]. The key features of the approach is to take a Bayesian approach with partial observations with the main goal of inferring the parameters in the model, making use of the fact that the likelihood is tractable. Example calculations confirm the computational efficiency of these methods.

David Campbell gave a presentation starting with a case study in parameter estimation from the Dow Chemical company in 1981 with serious identifiability problems. With a poorly defined relationship between data and potential model structures Dave proposed several methods to learn about the problem by relaxing the proposed model structure, using the data to learn about how best to model the data and provide quantitative information about when parameters are identifiable or not [16].

Alexandre Bouchard-Cote spoke about difficulties with high-dimensional problems by focusing on well-known ODE models and MCMC Bayesian inference, discussing case studies of benchmarking MCMC methods, most of which fail except for a technique called parallel tempering, which is a novel method with the input of a Markov Chain where the output is a higher performance Markov Chain [3]. Examples, observations and implementation rules of thumb follow.

Thomas Prescott presented recent work that takes a multi-fidelity approach to approximate Bayesian computation for stochastic models of biochemical processes [15]. The key advance in the presented method is the use of a low-fidelity model to try and make an early accept or reject decision for a given parameter sample from the posterior. In the context that simulation of the high-fidelity model is instead required for this decision, the use of a common noise input to correlate output from the low- and high-fidelity models (using ideas from the multi-level Monte Carlo literature) allows for computational savings.

Ramon Grima gave a presentation motivating the kinds of data of interest reflecting single cell level signals of underlying gene networks, focusing on temporal snapshots so that measurements of temporal moments is the key quantity of interest. Grima shows that the posterior mode using moment-based inference methods is fast and accurate, and some analysis comparing the approximate results and exact results suggests a form of the systematic error in the likelihood approximation [9].

Simon Cotter gave a presentation about accelerated importance sampling to deal with MCMC sampling that are curved and thin ("banana shaped") and that often arise when variables cannot be easily observed. Without good ideas of the manifold, issues arise in standard methods, such as Metropolis Hastings or Importance Sampling. The idea is to work with Parallel Adaptive Importance Sampling (PAIS) and Transport Maps to Importance sampling, these ideas borrowed from applying Transport Maps to MCMC algorithms [5]. In general the transport maps are multidimensional Gaussians that can be sampled very simply. Implementations are discussed.

Barbel Finkenstädt Motivated her work by considering observations of a circadian oscillator with a delay and explores challenges associated with experimental measurements of such oscillators and demonstrates how to use an adaptive MCMC algorithms to estimate model parameters and then outlines methodologies that might be relevant to exploring spatial structure and spatial patterns [19].

4 Scientific Progress Made

Daily discussion sessions were used to isolate key problems / challenges faced by a range of the attendees, and to attempt to make progress towards tackling them. We focussed on a different problem, and used a different format each day, to stimulate discussion, new ideas and collaborations.

4.1 Day 1 discussion – "Burning questions"?

The meeting was perhaps fairly unique in the sense that it brought together researchers from a range of different application fields, and so no attendee was familiar with all others in advance of the meeting. As such, the meeting began with an introductory session, where participants introduced themselves, their research interests, and a "Burning question for the workshop". These questions provided the basis for the group discussions during the rest of the week. A sample of the points raised during this session are listed below.

- What can the process of model inference in biology learn from other disciplines? What makes biology unique, and where can biology borrow?
- We have two versions of the same model: a deterministic parametric model, data and inference; or a stochastic parametric model, internal noise, data, inference. What are sensible criteria for model choice?
- When selecting between models encoding different hypotheses, how do we cope with all our models being 'wrong'?
- How can / should we learn from imperfect models?
- Is there a role for benchmark data(/model) analysis tasks in this field?
- How can we leverage machine learning tools for inference?
- How can mathematical models be applied clinically to directly inform treatments for individual patients?
- How can we construct fast and accurate inference methods for single-cell data?
- How do we combine modern machine-learning approaches with mechanistic models?
- When, and how much, should we care about identifiability for biological models?
- How can we develop appropriate inference schemes for models that cross multiple scales?
- Which classes of model assumption can be assessed? Can assumption impact be measured / quantified?
- What are the best ways to balance objectives in optimization problems with mixed types of data?

The questions outlined above were the subject of many talks, and of subsequent discussions (both formal and informal). **Importantly, we are aware that such discussions between participants are continuing well after the workshop, and these discussions have the potential to lead to new scientific discoveries and the establishment of new scientific partnerships that would not have been created without the workshop**.

4.2 Day 2 discussion – "What should Gary do"?

Within his talk on Day 1, Gary Mirams posed lots of questions relevant to the workshop participants, centered around model development, parameter identifiability and optimal experimental design in the context of ion channel modelling. As a basis for concrete discussions, the Day 2 discussion involved participants working in small groups to suggest avenues for Gary to explore. Themes and ideas relevant to a range of projects that came out of the discussion were

- Would using a correlated error structure be appropriate? To what extent is heteroskedasticity an issue? Can cross validation help in exploring the noise structure?
- Can input signals be designed and / or modified to discriminate between different models (i.e. is model-guided experimental design possible)? Can the number of possible models be reduced iteratively by variation of the input signal? When should we use the simplest adequate model in place of the "best" model?
- Can we learn anything by attempting to fit a mixture of all postulated models? Can machine learning help with model design?
- What metrics are useful in comparing models and data?

4.3 Day 3 discussion – "Best practices for reporting"

The discussion on day 3 entered around best practices for reporting. The discussion involved participants working in small groups to identify the current practices, and how those could be improved in the future. The discussions were divided into five areas:

Benchmark experimental data sets. In field such as image analysis and machine learning, the availability of experimental data sets that can be used to test new algorithms has been vital to progress in the field. However, no such data sets exist for problems in parameter inference and model selection. Issues that were discussed in this respect included the lack of a repository where authors could submit data, access to high quality data (with a clear description of any pre-processing steps taken), whether different models and implementations would provide similar results. A recommendation going forward is that authors should post full data sets (wherever possible) within a publicly accessible repository, and provide a link to it within any published research article.

Bench-marking problems. Discussions were centered around the different types of problems for which algorithm bench-marking is necessary, including those associated with parameter estimation, model selection, approximations and models. The creation of synthetic data sets was recommended, along with the use of modular approach for different aspects of modelling and inference.

Software best practices. With increasingly sophisticated models, data sets and inference approaches now routinely in use, the use of software best practices is vital, yet not generally adopted. The group recommends that all software should be modular, well-documented and developed using version control.

Repositories / software packages / Journals. One issue raised is the current lack of an obvious "home" for papers in this newly emerging field, and the potential for a Special Issue was discussed as a means to raise the profile of the field. Github was widely viewed as a good platform for hosting software / code arising from research projects, but the group suggested that code / software should also be included within publications (and deposited in a journal repository, as is currently the case for Supplementary Material).

Publication best practices. The group recommended publishing at least pseudo-code to accompany (where relevant) all research papers, and that any code submitted with a manuscript should also be peer-reviewed. They recognised that there needs to be a greater awareness of the contributions of researchers in developing code / software, and that all code should be published with papers, as should data (experimental and synthetic).

The overwhelming point that came out of the Day 3 discussion, was the importance of building a new research community for the field, to promote collaboration and the sharing of ideas and best practices.

4.4 Day 5 discussion – "All models are priors" debate

The discussion on Day 4 took the form of a lively debate over whether "All models are priors". This raised some interesting points over how to approach modelling and inference!

5 Outcome of the Meeting

The meeting has seeded a new community in the broad field of quantitative approaches to biology and ecology, that is focussed on connecting models and quantitative data using statistical techniques. Going forward, it will be important to maintain this momentum, by having a similar meeting perhaps every 2-3 years that brings together the community to share and recent results and developments, and establish new collaborations, by having small focussed meetings to work on key problems, and by putting together a special issue for publication in an interdisciplinary journal (such as *Journal of the Royal Society Interface Focus* or *Bulletin of Mathematical Biology*).

6 Participant response

Alex Browning. I am a fairly new PhD student, who has only undertaken a few research projects in the past. The BIRS workshop gave me a fantastic opportunity to present and discuss my work and ideas with a large number of new connections that I made. I believe it aided me enormously in terms of future prospects, as well as future research ideas.

Simon Cotter. It was one of the most productive and gratifying workshops that I have ever been on. The scientific programme was superb, and the organisation by everyone at BIRS was top-notch. It was also a stunning location and I thoroughly enjoyed my trip. In short, it was well worth the long journey, and I hope that I will get a chance to visit again in the future.

John Fricks. The BIRS workshop was especially helpful in providing insights on which computational frameworks and tools to use in order to move my research forward. I made a lot of great contacts and met potential collaborators with whom I expect to keep in contact. One of the important takeaways from the meeting was a plan for this to be a first step in building a broader community.

Priscilla Greenwood. This was an outstanding workshop for me in terms of making important new contacts: Adelle Coster from Sydney, whom I was able to help with a problem she presented in her talk, also had time to get acquainted and plan to meet again. In addition, after my talk, we found many points of common interest with Ramon Grima from Edinburgh, and discussed for several hours. We have already exchanged several of our papers and also other references. This contact may well lead to joint work. Will keep you posted. The workshop, in general, was excellent. It brought together people who had several different approaches (a number centered around approximate Bayesian computation, a rather new version of MCMC which reduces computing time in complex models with much or little data) to inference for ODE, SDE and SPDE-type models arising in all sorts of math-bio settings, cell biology (very important current topic) and epidemiology, various medical applications, and so on. Most of the participants, mostly from UK, North America and Australia-New Zealand, had not met. So the group tended to assemble as a whole, rather than

breaking into small circles of old buddies as sometimes happens. I felt that I left with lots of new friends! An outstanding contributor was one of the organizers, our own Dan Coombs, who set a number of discussion topics which drew lively, inclusive, sometimes heated discussion, one of these sessions each day. I was lucky and very much pleased to be part of this workshop!

References

[1] R. E. Baker and M. J. Simpson, Correcting mean-field approximations for birth-death-movement processes. *Phys. Rev. E* **82** (2010), 041905.

[2] R. N. Binny, M. J. Plank and A. James, Spatial moment dynamics for collective cell movement incorporating a neighbour-dependent directional bias. *J. Roy. Soc. Interface* **12** (2015), 20150228.

[3] A. Bouchard-Coté, S. J. Vollmer and A. Doucet, The Bouncy Particle Sampler: A non-reversible rejection-free Markov chain Monte Carlo method. *J. Am. Stat. Assoc.* **113** (2018), 855–867.

[4] A. P. Browning, P. Haridas and M. J. Simpson, A Bayesian sequential learning framework to parametrise continuum models of melanoma invasion into human skin. *Bull. Math. Biol.* **81** (2018), 676–698.

[5] S. L. Cotter, I. G. Kevrekidis and P. Russell, Transport map accelerated adaptive importance sampling, and application to inverse problems arising from multiscale stochastic reaction networks. *arXiv* (2018), 1901.11269.

[6] C. W. Gary and A. C. F. Coster, A receptor state space model of the insulin signalling system in glucose transport. *Math. Med. Biol.* **32** (2015), 457–473.

[7] M. Chung and E. Haber, Experimental design for biological systems. *SIAM J. Control Optim.* **50** (2012), 471–489.

[8] M. Chung, M. Binois, R.B. Gramacy, D.J. Moquin, A.P. Smith and A.M. Smith, Parameter and uncertainty estimation for dynamical systems using surrogate stochastic processes. *arXiv* (2018).

[9] F. Frölive, P. Thomas, A Kazeroonian, F. J. Theis, R. Grima and J. Hasenauer, Inference for stochastic chemical kinetics using moment equations and system size expansion. *PLoS Comput. Biol.* **12** (2016), e1005030.

[10] R. Johnstone, R. Bardenet, L. Polonchuk, M. Davies, D. Gavaghan and G. Mirams, Hierarchical Bayesian fitting of concentration-effect models to ion channel screening data. *J. Pharmacy. Toxicol. Methods* **88** (2017), 198.

[11] B. Lambert, A. L. MacLean, A. G. Fletcher, A. N. Combes, M. H. Little and H. M. Byrne, Bayesian inference of agent-based models: a tool for studying kidney branching morphogenesis. *J. Math. Biol.* **76** (2018), 1673–1697.

[12] J. Liepe, S. Filippi, M. Komorowski and M. P. H. Stumpf, Maximizing the information content of experiments in systems biology. *PLoS Comput. Biol.* **9** (2013), e1002888.

[13] A. M. Middleton, C. Fleck and R. Grima, A continuum approximation to an off-lattice individual-cell based model of cell migration and adhesion. *J. There. Biol.* **10** (2014), 220–232.

[14] G. Pokharel and R. Deardon, Gaussian process emulators for spatial individual?level models of infectious disease. *Can. J. Stat.* **44** (2016), 480–501.

[15] T. P. Prescott and R. E. Baker, Multifidelity approximate Bayesian computation. *arXiv* (2018), 1811.09550.

[16] J. O. Ramsay, G. Hooker, D. Campbell and J. Cao, Parameter estimation for differential equations: a generalized smoothing approach. *J. Roy. Stat. Soc. Series B* **69** (2007), 741–796.

[17] T. Ryder, A. Golightly, A. S. McGough and D. Prangle, Black-box autoregressive density estimation for state-space models. *arXiv preprint* 1811.08337.

[18] D. Silk, P. D. Kirk, C. P. Barnes, T. Toni and M. P. H. Stumpf, Model selection in systems biology depends on experimental design. *PLoS Comput Biol* **10** (2014), e1003650.

[19] S. Tiberi, M. Walsh, M. Cavallaro, D. Hebenstreit and B. Finkenstädt, Bayesian inference on stochastic gene transcription from flow cytometry data. *Bioinformatics* **34** (2018), i647–655.

[20] T. Toni, D. Welch, N. Strelkowa, A. Ipsen and M. P. H. Stumpf, Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. Roy. Soc. Interface* **6** (2009), 187–202.

[21] D. J. Wilkinson, *Stochastic modelling for systems biology*. Chapman & Hall/CRC, 2018.