

Preface to the third edition

I was fairly pleased with the way that the second edition of this book, published in 2011, turned out. In particular, the incorporation of the **R** package, `smfSB`, proved to be a very popular development, improving the accessibility and reproducibility of the algorithms described. With this third edition I have not substantially re-written the existing text, but instead concentrated on incorporating new material which covers some important topics missing from the second edition, together with appropriately extending the software and accompanying documentation associated with the book.

The first development relates to associated software, which has little impact on the text, but has potentially significant impact on the way that the book is used for both teaching and research. For example, I've produced a small software library for reading SBML models into **R** for subsequent analysis and simulation using the `smfSB` **R** package. So it is now possible to directly use SBML models (designed with discrete stochastic simulation in mind) as the basis for simulation experiments. This was a fairly obvious omission from the second edition of this book, but at the time of producing the second edition the software libraries required to facilitate this (`libSBML` **R** bindings) were too immature to make this practical. There are other improvements and developments relating to software associated with the book, and these are now all discussed in a new Appendix. Further information is provided about tools for working with the 'SBML shorthand' notation I use in the text, and there is a description of a new software library associated with the text. This library, written in a fast, efficient, compiled language, Scala, replicates most of the functionality associated with the **R** packages, and therefore provides a convenient route for people wanting to move beyond working in (slow) dynamically typed languages like **R** to better, more efficient, general-purpose languages. This will probably be more relevant to researchers than students, but should prove to be very useful for people needing to make that transition.

The second significant addition is the inclusion of a new chapter on spatially extended systems (stochastic reaction-diffusion models). Again, this was a fairly obvious omission from the second edition, but for that edition I really wanted to concentrate on ensuring that there was adequate coverage of the well-mixed case. Spatial stochastic modelling is a huge topic, and this new chapter barely scratches the surface, but some of the most important topics are now covered, and code is included that will allow readers to easily take a well-mixed stochastic biochemical reaction network model and investigate its behaviour as a reaction-diffusion system in 1 and 2 dimensions.

The third major addition concerns the chapter on inference for stochastic kinetic models. An important addition for the second edition was discussion of likelihood-

free methods and particle MCMC, including some example code. These sections have been significantly updated for the third edition, and new sections have been added covering the use of approximate Bayesian computation (ABC) methods for parameter inference, including sequential ABC methods, and associated code. This reflects the increasing use of ABC methods in systems biology in recent years.

There are also numerous other minor updates throughout. Hopefully these significant new additions to the text and the associated software ecosystem improve the range of courses for which this text is applicable.

Darren J. Wilkinson
Newcastle upon Tyne

Preface to the second edition

I was keen to write a second edition of this book even before the first edition was published in the spring of 2006. The first edition was written during the latter half of 2004 and the first half of 2005 when the use of stochastic modelling within computational systems biology was still very much in its infancy. Based on an inter-disciplinary masters course I was teaching, I saw an urgent need for an introductory textbook in this area, and tried in the first edition to lay down all of the key ingredients needed to get started. I think that I largely succeeded, but the emphasis there was very much on the 'bare essentials' and accessibility to non-mathematical readers, and my goal was to get the book published in a timely fashion, in order to help advance the field. I would like to think that the first edition of this text has played a small role in helping to make stochastic modelling a much more mainstream part of computational systems biology today. But naturally there were many limitations of the first edition. There were several places where I would have liked to have elaborated further, providing additional details likely to be of interest to the more mathematically or statistically inclined reader. Also, the latter chapters on inference from data were rather limited and lacking in concrete examples. This was partly due to the fact that the whole area of inference for stochastic kinetic models was just developing, and so it wasn't possible to give a coherent overview of the problem from an introductory viewpoint. Since publishing the first edition there have been many interesting developments in the use of 'likelihood-free' methods of Bayesian inference for complex stochastic models, and so the latter chapters have now been re-written to reflect this more modern perspective, including a detailed case study accompanied by working code examples.

Of course the whole field has moved on considerably since 2005, and so the second edition is also an opportunity to revise and update, and to change the emphasis of the text slightly. The Systems Biology Markup Language (SBML) has continued to evolve, and SBML Level 3 is now finalised. Consequently, I have updated all of the examples to Level 3, which is likely to remain the standard encoding for dynamic biological models for the foreseeable future. I have also taken the opportunity to revise and update the R code examples associated with the book, and to bundle them all together as an R package (`smf.sb`). This should make it much easier for people to try out the examples given in the book. I have also re-written and re-structured all of the code relating to simulation, analysis, and inference for stochastic kinetic models. The code is now structured in a more modular way (using a functional programming style), making it easy to 'bolt together' different models, simulation algorithms, and analysis tools. I've created a new website specific to this second edition, where I will

keep links, resources, an errata, and up-to-date information on installation and use of the associated R package.

The new edition contains more background material on the theory of Markov processes and stochastic differential equations, providing more substance for mathematically inclined readers. This allows discussion of some of the more advanced concepts relating to stochastic kinetic models, such as random time-change representations, Kolmogorov equations, Fokker–Planck equations and the linear noise approximation. It also enables simple modelling of ‘extrinsic’ in addition to ‘intrinsic’ noise. This should make the text suitable for use in a greater range of courses. Naturally, in keeping with the spirit of the first edition, all of the new theory is presented in a very informal and intuitive way, in order to keep the text accessible to the widest possible readership. This is not a rigorous text on the theory of Markov processes (there are plenty of other good texts in that vein) — the book is still intended for use in courses for students with a life sciences background.

I’ve also updated the references, and provided new pointers to recent publications in the literature where this is especially pertinent. However, it should be emphasised that the book is not intended to provide a comprehensive survey of the stochastic systems biology literature — I don’t think that is necessary (or even helpful) for an introductory textbook, and I hope that people working in this area accept this if I fail to cite their work.

So here it is, the second edition, completed at last. I hope that this text continues to serve as an effective introduction to the area of stochastic modelling in computational systems biology, and that this new edition adds additional mathematical detail and computational methods which will provide a stronger foundation for the development of more advanced courses in stochastic biological modelling.

Darren J. Wilkinson

Newcastle upon Tyne, June 2011

Preface to the first edition

Stochastic models for chemical and biochemical reactions have been around for a long time. The standard algorithm for simulating the dynamics of such processes on a computer (the 'Gillespie algorithm') was published nearly 30 years ago (and most of the relevant theory was sorted out long before that). In the meantime there have been dozens of papers published on stochastic kinetics, and several books on stochastic processes in physics, chemistry, and biology. Biological modelling and biochemical kinetic modelling have been around even longer. These distinct subjects have started to merge in recent years as technology has begun to give real insight into intra-cellular processes. Improvements in experimental technology are enabling quantitative real-time imaging of expression at the single-cell level, and improvement in computing technology is allowing modelling and stochastic simulation of such systems at levels of detail previously impossible. The message that keeps being repeated is that the kinetics of biological processes at the intra-cellular level are stochastic, and that cellular function cannot be properly understood without building that stochasticity into *in silico* models. It was this message that first interested me in systems biology and in the many challenging statistical problems that follow naturally from this observation.

It was only when I came to try and teach this interesting view of computational systems biology to graduate students that I realised there was no satisfactory text on which to base the course. The papers assumed far too much background knowledge, the standard biological texts didn't cover stochastic modelling, and the stochastic processes texts were too far removed from the practical applications of systems biology, paying little attention to stochastic biochemical kinetics. Where stochastic models do crop up in the mainstream systems biology literature, they tend to be treated as an add-on or after-thought, in a slightly superficial way. As a statistician I see this as problematic. The stochastic processes formalism provides a beautiful, elegant, and coherent foundation for chemical kinetics, and there is a wealth of associated theory every bit as powerful and elegant as that for conventional continuous deterministic models. Given the increasing importance of stochastic models in systems biology, I thought it would be particularly appropriate to write an introductory text in this area from this perspective.

This book assumes a basic familiarity with what might be termed high school mathematics. That is, a basic familiarity with algebra and calculus. It is also helpful to have had some exposure to linear algebra and matrix theory, but not vital. Since the teaching of probability and statistics at school is fairly patchy, essentially nothing will be assumed, though obviously a good background in this area will be very helpful. Starting from here, the book covers everything that is necessary for a good

appreciation of stochastic kinetic modelling of biological networks in the systems biology context. There is an emphasis on the necessary probabilistic and stochastic methods, but the theory is rooted in the intended application, and no time is wasted covering interesting theory that is not necessary for stochastic-kinetic modelling. On the other hand, more-or-less everything that is necessary is covered, and the text (at least up to Chapter 8) is intended to be self-contained. The final chapters are necessarily a little more technical in nature, as they concern the difficult problem of inference for stochastic kinetic models from experimental data. This is still an active research area, and so the main aim here is to give pointers to the existing literature and provide enough background information to render that literature more accessible to the non-specialist.

The decision to make the book practically oriented necessitated some technological choices that will not suit everyone. The two key technologies chosen for illustrating the theory in this book are SBML and R. I hope that the choice of the Systems Biology Markup Language (SBML) for model representation is not too controversial. It is the closest thing to a standard that exists in the systems biology area, and there are dozens of software tools that support it. Of course, most people using SBML are using it to encode continuous deterministic models. However, SBML Level 2 and beyond are perfectly capable of encoding discrete stochastic models, and so one of the reasons for using it in this text is to provide some working examples of SBML models constructed with discrete stochastic simulation in mind.

The other technological choice was the use of the statistical programming language R. This is likely to be more controversial, as there are plenty of other languages that could have been used. It seemed to me that in the context of a textbook, using a very high-level language was most appropriate. In the context of stochastic modelling, a language with good built-in mathematical and statistical support also seemed highly desirable. R stood out as being the best choice in terms of built-in language support for stochastic simulation and statistical analysis. It also has the great advantage over some of the other possible choices of being completely free open-source software, and therefore available to anyone reading the book. In addition, R is being used increasingly in bioinformatics and other areas of systems biology, so hopefully for this reason too it will be regarded as a positive choice.

This book is intended for a variety of audiences (advanced undergraduates, graduate students, postdocs, and academics from a variety of backgrounds), and exactly how it is read will depend on the context. The book is certainly suitable for a variety of graduate programs in computational biology. I will be using it as a text for a second-semester course on a masters in bioinformatics programme that will cover much of Chapters 1, 2, 4, 5, 6, and 7 (most of the material from Chapter 3 will be covered in a first-semester course). However, the book has also been written with self-study in mind, and here it is intended that the entire book be read in sequence, with some chapters skipped depending on background knowledge. It is intended to be suitable for computational systems biologists from a continuous deterministic background who would like to know more about the stochastic approach, as well as for statisticians who are interested in learning more about systems biology (though statisticians will probably want to skip most of Chapters 3, 4, and 5). It is worth pointing

out that Chapters 9 and 10 will be easier to read with a reasonable background in probability and statistics. Though it should be possible to read these chapters without such a background and still appreciate the key concepts and ideas, some of the technical details may be difficult to understand fully from an elementary viewpoint.

Writing this book has been more effort than I anticipated, and there were one or two moments when I doubted the wisdom of taking the project on. On the whole, however, it has been an interesting and rewarding experience for me, and I am pleased with the result. I know it is a book that *I* will find useful and will refer to often, as it integrates a fairly diverse literature into a single convenient and notationally consistent source. I can only hope that others share the same view, and that the book will help make a stochastic approach to computational systems biology more widely appreciated.

Darren J. Wilkinson
Newcastle upon Tyne, October 2005