MODELLING

Stochastic modelling for quantitative description of heterogeneous biological systems

Darren J. Wilkinson

Abstract | Two related developments are currently changing traditional approaches to computational systems biology modelling. First, stochastic models are being used increasingly in preference to deterministic models to describe biochemical network dynamics at the single-cell level. Second, sophisticated statistical methods and algorithms are being used to fit both deterministic and stochastic models to time course and other experimental data. Both frameworks are needed to adequately describe observed noise, variability and heterogeneity of biological systems over a range of scales of biological organization.

Continuous deterministic mathematical model
A model that does not contain any element of unpredictability, and that describes the smooth and gradual change of model elements (such as biochemical substances) according to pre-determined mathematical rules. The precise behaviour of the model is entirely pre-determined (and hence, in principle, predictable) from the form of the equations and the starting conditions.

School of Mathematics & Statistics and the Centre for Integrated Systems Biology of Ageing and Nutrition (CISBAN), Newcastle University, Newcastle upon Tyne, Tyne and Wear NE1 7RU, UK.
e-mail: d.j.wilkinson@ncl.ac.uk
doi:10.1038/nrg2509
Published online
13 January 2009

Systems biology[1] aims to move beyond the study of single biomolecules and the interaction between specific pairs of molecules; its goal is to describe, in quantitative terms, the dynamic systems behaviour of complex biological systems that involve the interaction of many components. Traditional reductionist genetic and molecular biology approaches have yielded huge amounts of data, but understanding how low-level biological data translates into functioning cells, tissues and organisms remains largely elusive. Now that life scientists possess an extensive 'parts list' for biology, we can begin to think about how the function of a biological system arises from dynamic interactions between its parts. As even simple dynamic systems can exhibit a range of complex behaviour, such an approach requires quantitative mathematical and statistical modelling of biological system dynamics. At the level of cellular modelling, this ideally requires time course data on the abundance of many different biomolecules at single-cell resolution.

Traditionally, systems dynamics have been described by using continuous deterministic mathematical models. However, it has recently been acknowledged that biochemical kinetics at the single-cell level are intrinsically stochastic[2]. It is now generally accepted that stochastic models are necessary to properly capture the multiple sources of heterogeneity needed for modelling biosystems in a realistic way. However, such models come at a price; they are computationally more demanding than deterministic models, and considerably more difficult to fit to experimental data.

Statistics is the science concerned with linking models to data, and as such it is absolutely pivotal to the success of the systems biology vision. Statistical approaches to inferring the parameters of deterministic and stochastic biosystems models provide the best way to extract maximal information from biological data. Effective methods for statistically estimating stochastic models by using time course data have only recently appeared in the systems biology literature; these techniques are the final piece of the puzzle needed to describe biological dynamics in a quantitative framework.

This article reviews the key issues that need to be understood to describe biological heterogeneity properly, the approaches that have been used and the range of problems that they solve, together with the most promising avenues for further development. Many of the examples in the literature concern single-celled organisms such as bacteria and yeast; however, heterogeneity is present in all biological systems, and separating intrinsic stochasticity from genetic and environmental sources[3] is likely to become increasingly important in the context of human genetics and complex diseases in the near future.

Basic modelling concepts: a working example

One of the principal aims of systems biology is to test whether our understanding of a complex biological process is consistent with observed experimental data. As dynamic systems exhibit complex behaviour, our understanding must be encoded in quantitative mathematical models. A lack of consistency between the model and the data indicates that further research is required to

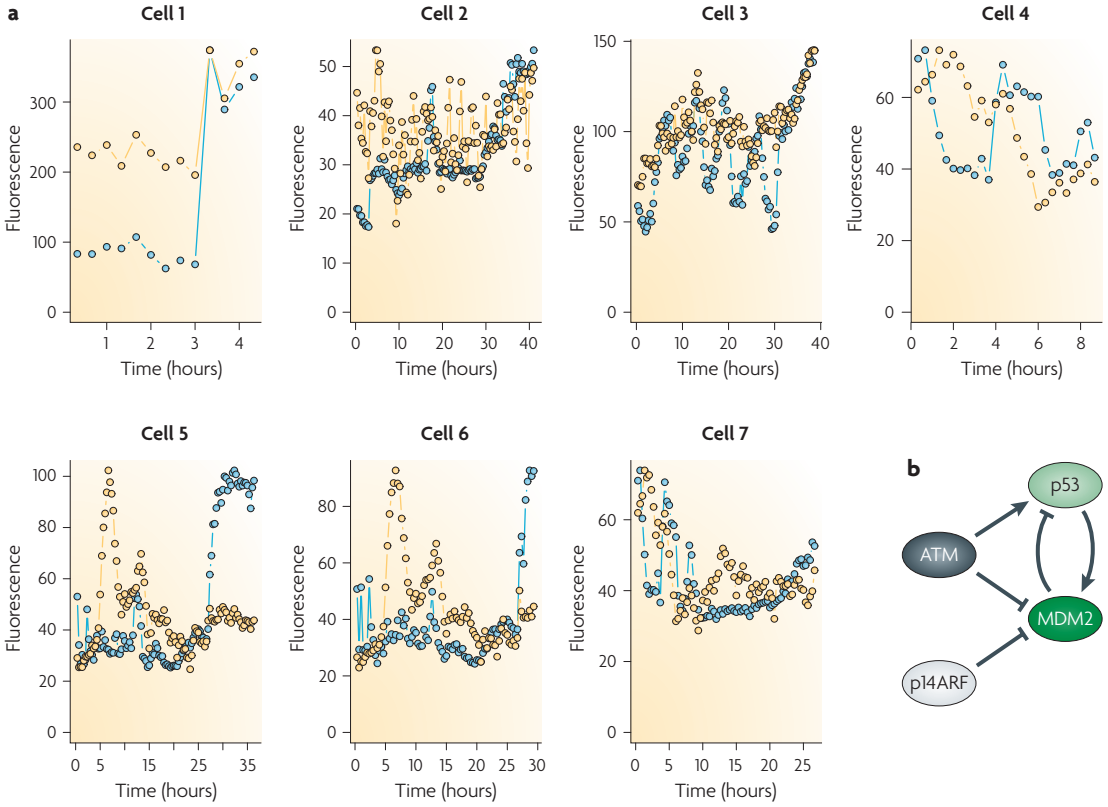


Figure 1 | **Fluctuations in p53 and MDM2 levels in single cells. a** | Image analysis can be used to extract time courses of expression levels from time-lapse microscope movies. The plots show the measured fluorescence levels for seven individual cells from one particular movie (movie 2 in data from REF. 7, provided by the authors); the tumour suppressor protein p53 is represented by blue circles, and the ubiquitin E3 ligase MDM2 is represented by yellow circles. Although there is some evidence of p53 and MDM2 oscillations, there is clearly a highly heterogeneous cellular response. **b** | The essential interactions between p53, MDM2, and key signalling molecules ataxia telangiectasia mutated (ATM) and the cyclin-dependent kinase inhibitor p14ARF (also known as CDKN2A). p53 activates transcription of MDM2. MDM2 then binds to p53, thereby enhancing its degradation[8,9]. p53 can be activated by the kinase ATM, which is activated by DNA damage; ATM phosphorylates p53 and MDM2, this prevents the binding of p53 to MDM2 and enhances MDM2 degradation, thereby allowing accumulation of active p53. MDM2 can also be inactivated by p14ARF.

complete our understanding of the system under study. Consistent models can be used to make further testable predictions for more independent validation, and also to carry out *in silico* investigation of the system behaviour that would be difficult or time consuming to do entirely in the laboratory. These concepts can be illustrated using the example of oscillations and variability in the well-characterized p53–MDM2 system.

The human tumour suppressor protein p53 (encoded by the *TP53* gene) is a transcription factor that has an important role in regulating the cell cycle, tumour suppression and DNA damage response[4]. Population level data showed only a single peak in p53 expression, followed by decay back to basal levels. More recently, however, single cell assays in MCF7 breast cancer cell lines have revealed that levels of p53 sometimes seem to oscillate in response to radiation-induced DNA damage[5–7]. For example, the Alon laboratory measured p53 and MDM2 levels in single cells over time using two fluorescent reporters[7]. FIGURE 1a shows clearly a highly heterogeneous cellular response despite some evidence of p53 and MDM2 oscillations.

Oscillations are indicative of negative feedback in the system dynamics. We would therefore like to understand the underlying mechanisms, and to test that understanding by developing quantitative and predictive models of the system behaviour. The essential feedback feature of this system is well known: p53 activates transcription of MDM2, a ubiquitin E3 ligase, which in turn binds to p53 and thereby enhances its degradation[8,9]. The signal for p53 activation can come from more than one source. In MCF7 cell lines, which do not express the cyclin-dependent kinase inhibitor p14ARF (also known as CDKN2A), the strongest signal probably comes from the kinase ATM (ataxia telangiectasia mutated), which is activated by DNA damage; ATM phosphorylates both p53 and MDM2, blocking their binding to each other and enhancing MDM2 degradation, thereby allowing accumulation of active p53 (FIG. 1b).

Many systems biology models are concerned with intracellular processes, and therefore operate (conceptually, at least) at the level of a single cell. Most stochastic and deterministic models for chemical reaction network kinetics make the assumption that cellular compartments

Stochastic model
A model that contains an element of unpredictability or randomness specified in a precise mathematical way. Each run of a given model will produce different results, but the statistical properties of the results of many such runs are pre-determined by the mathematical formulation of the model.

can be regarded as small well-stirred containers, thus ignoring spatial effects, and describe the dynamics of the process of interest using a set of biochemical reactions[10]. The differences between stochastic and deterministic approaches relate to the assumptions made regarding the nature of the kinetic processes associated with the reaction network.

Deterministic models

The classical approach to chemical kinetics is to assume that reactants are abundant and have a level measured on a continuous scale, traditionally in units of concentration. In the p53 example, reactants will be proteins and complexes such as p53, MDM2, p53–MDM2, phosphorylated p53, as well as the mRNA molecules that encode the different proteins. The state of the system at any particular instant is therefore regarded as a vector (or list) of amounts or concentrations. Furthermore, the changes in amount or concentration are assumed to occur by a continuous and deterministic process. The velocity of each reaction is specified using a rate equation that typically assumes mass action kinetics or is based on an enzyme kinetic law (such as Michaelis–Menten or Hill kinetics)[11]. The way in which the state of the system evolves can be described mathematically (by using ordinary differential equations (BOX 1)). In certain simple but usually not biologically realistic cases, these equations can be solved to give an explicit formula that describes the time course trajectory. In more complicated scenarios, such as the p53 example described above, computational methods are used that provide only approximate (but typically accurate) solutions.

Although some deterministic models of the p53–MDM2 system have been proposed in the literature[6,7,12,13], they are unsatisfactory for several reasons. The most fundamental limitation of deterministic models is that they inevitably fail to explain the highly noisy and heterogeneous observed cellular response to DNA damage. The obvious lack of agreement between the model and the data cannot be attributed to genetic or environmental effects, as these have been largely eliminated by the experimental design. It is therefore difficult to make any sensible assessment of the extent to which such models explain the observed data. Another limitation of deterministic models is that they do not span multiple scales. Either the model oscillates (as suggested by the single-cell assays), or it has a single peak in p53 expression (as observed in population level data). It is difficult to reconcile these two observations without accepting a heterogeneous cellular response, and in practice this involves introducing stochasticity into the models. By contrast, an essentially stochastic model based on the known biochemical mechanisms has recently been described[14]. This simple model shows that the heterogeneity observed in the experimental data is entirely consistent with intrinsic stochasticity in the system; it also has the property that the population average of the p53 levels of many single cells over time has the observed single peak in p53 expression.

Modellers aim to find simple explanations for a range of complex and sometimes apparently contradictory experimental observations. Here, a single simple mechanistic model simultaneously explains how p53 levels can oscillate and why they do not oscillate in some cells, the origins of stochasticity and heterogeneity in the cellular response, and the apparent conflict between the single-cell data and population level data. No comparatively simple deterministic model can do this. Furthermore, because the stochastic model exhibits a similar range of behaviour to the experimental data, it becomes meaningful to try and make a serious assessment of how well such a model matches the experimental data, and to try and use the experimental observations to improve our knowledge about the model parameters[15].

Stochastic models

The continuous deterministic approach to modelling biochemical reaction networks fails to capture many important details of a biological process and the experimental data that relates to this process. The 'missing detail' manifests itself as a degree of apparent unpredictability of the system. As a result, the single-cell dynamics of biological systems seem noisy, or stochastic, with these terms being used more or less interchangeably. Heterogeneity is then a phenotypic consequence for a cell population given stochastic single-cell dynamics. Stochasticity and heterogeneity are aspects of model biological system behaviour that cannot be ignored, and attempts to refine experimental techniques to eliminate them are both hopeless and misguided[2,16,17].

There are multiple sources of stochasticity and heterogeneity in biological systems, and these can, and often do, have important consequences for understanding overall system behaviour. Stochasticity influences genetic selection and evolution[18,19]; biological systems have also developed strategies for both exploiting[20] and suppressing[18] biological noise and heterogeneity[21]. Any useful predictive model of the system must therefore account for a degree of intrinsic unpredictability.

The only satisfactory quantitative modelling framework that takes into account the inherent unpredictability of a system is based on probability theory. Statistical mechanical arguments are used to understand the probabilistic behaviour of the dynamic stochastic process associated with the biochemical network. The dynamics of a biological system can be modelled using a Markov jump process, whereby any change in the system occurs discretely after a random time period, with the change and the time both depending only on the previous state[22–25]. This is a well-understood model from the theory of stochastic processes. It has been known for decades that this framework can be applied to the simulation of stochastic chemical kinetics[26], but it did not become a well-established approach in biology until the late 1990s[27], when experimental techniques became precise enough to show that experimental findings could be modelled accurately only by incorporating stochasticity[28]. Stochastic modelling has a long tradition in other areas of biological modelling, including population dynamics[24,29].

Michaelis–Menten
A simple kinetic law that modifies the rate of conversion from substrate to product based on enzyme concentration.

Hill kinetics
A more complex enzyme kinetic law than simple Michaelis–Menten kinetics.

Ordinary differential equation
A mathematical equation involving differential calculus. In simple cases, explicit formulas can be derived for their solution, but typically they must be numerically integrated on a computer.

Probability theory
The mathematical theory of chance, randomness, uncertainty and stochasticity.

Markov jump process
A class of stochastic processes that is well studied in probability theory and that includes the class of processes described by stochastic chemical kinetics.

Stochastic chemical kinetics
A chemical kinetic theory which recognizes that molecules are discrete entities, and that reaction events occur at random when particular combinations of molecules interact.

Probability distribution
A precise mathematical description of a stochastic quantity.

As in the deterministic case, some simple network models are analytically tractable. In these simple situations, the full probability distribution for the state of the biological system over time can be calculated explicitly. However, as for the deterministic case, the class of solvable models is small, mainly covering those models that contain only single-molecule reactions. As almost all interesting systems involve interactions between molecules of different types, these simple models do not cover systems of genuine practical interest. Here, too,

Box 1 | **A simple model for protein production and degradation**

Consider the following artificial model for production and degradation of a single protein, X: the protein is produced at a constant rate α, and each protein molecule is independently degraded at a constant rate μ. This can be written using chemical reaction notation as:

$$\varnothing \xrightarrow{\alpha} X$$

$$X \xrightarrow{\mu X} \varnothing$$

Let the number of molecules at time t be denoted X_t, and assume that there are initially no protein molecules, so that $X_0 = 0$. The plots show the case $\alpha = 1$, $\mu = 0.1$. The parameters are purely illustrative and not intended to model any real biological system.

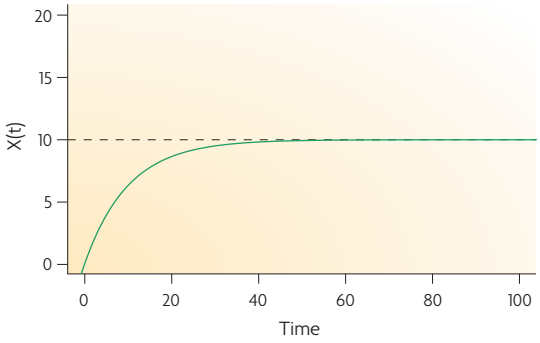
The first plot shows the standard reaction rate equation (RRE) model. Although in this case this model captures the essential 'shape' of the discrete stochastic (Markov jump process) model, shown in the second plot, it completely ignores the substantial variability. The final plot shows the chemical Langevin equation (CLE) model. Although this model sacrifices discreteness, it effectively captures both the shape and variability of the discrete stochastic solution, despite the low copy numbers involved. Note that although in the case of this simple model, the equilibrium means (and, indeed, time-varying means) of the stochastic models match the deterministic model, in general this is not the case. The plots for the stochastic models show a single realization of the process, based on independent noise processes.

Continuous deterministic model (RRE):

$$\frac{dX}{dt} = \alpha - \mu X$$

Solution: $X_t = \frac{\alpha}{\mu}(1 - e^{-\mu t})$

Equilibrium: $X_\infty = \alpha/\mu$



Discrete stochastic model:

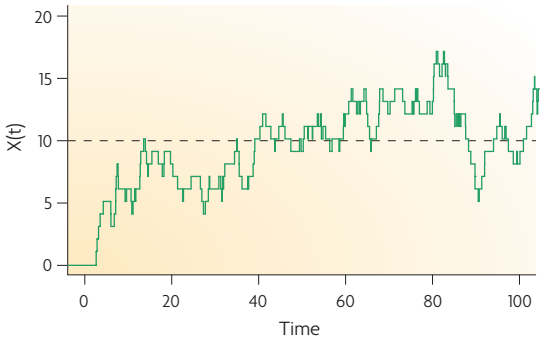
$$\Pr(X_{t+dt} = x + 1 \mid X_t = x) = \alpha\, dt$$

$$\Pr(X_{t+dt} = x - 1 \mid X_t = x) = \mu x\, dt$$

Solution: $X_t \sim \text{Poisson}\left(\frac{\alpha}{\mu}[1 - e^{-\mu t}]\right)$

Equilibrium distribution: $X_\infty \sim \text{Poisson}(\alpha/\mu)$

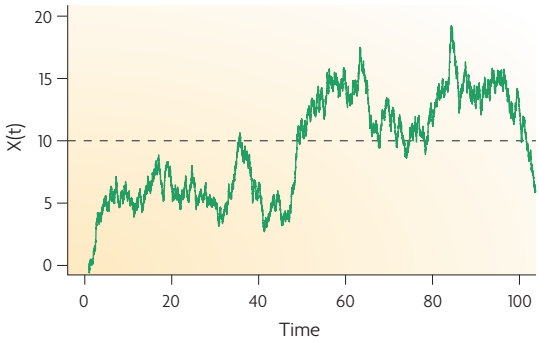
$$E(X_\infty) = \text{Var}(X_\infty) = \alpha/\mu$$



Continuous stochastic model (CLE):

$$dX_t = (\alpha - \mu X_t)dt + \sqrt{\alpha + \mu X_t}\, dW_t$$

At equilibrium: $E(X_\infty) = \text{Var}(X_\infty) = \alpha/\mu$



Box 2 | **Outline of the Gillespie algorithm**

The Gillespie algorithm is used to simulate stochastic time course trajectories of the state of a chemical reaction network. The essential structure of this discrete event simulation algorithm is outlined below.

- Step 1: set the initial number of molecules of each biochemical species in the reaction network and set the simulation time to zero.
- Step 2: on the basis of the current molecular abundances, calculate the propensity for each possible reaction event.
- Step 3: using the current propensities, simulate the time to the next reaction event, and update the simulation time accordingly (the larger the reaction propensities, the shorter the time to the next event).
- Step 4: pick a reaction event at random, with probabilities determined by the reaction propensities (higher propensities lead to higher probability of selection), and update the number of molecules accordingly.
- Step 5: record the new simulation time and state.
- Step 6: check the simulation time. If the simulation is not yet finished, return to step 2.

To give an explicit example, consider using the Gillespie algorithm to generate a realization from the simple model described in the discrete stochastic model of BOX 1, which considers the production and degradation of a molecule. At each point in the simulation, the time to the next reaction event is simulated (and the expected time to wait will be shorter the more molecules there are), and a decision will need to be made as to whether the reaction should be a synthesis or a degradation event, with the probability of degradation increasing as the number of molecules in the system increases. For an accessible introduction to the Gillespie algorithm, and stochastic modelling for systems biology more generally, see REF. 25.

numerical simulation of the process on a computer is the key tool used for understanding the system behaviour. For Markov jump process models, an algorithm known in this context as the stochastic simulation algorithm (but more commonly known as the Gillespie algorithm) is used to generate exact realizations (or 'runs') of the Markov jump process[26]. The algorithm generates time course trajectories of the system state over a given time window, starting from a given initial system state (BOX 2). Of course, these realizations are stochastic, and are therefore different for each run of the simulation model. They are 'exact' in the sense that each run is an independent realization from the true underlying process; properties deduced about the probabilistic nature of the process from multiple runs can be made arbitrarily accurate by averaging over a sufficient number of runs to reduce the Monte Carlo error associated with the estimates. Accessible introductions to methods of stochastic simulation for reaction networks can be found in REFS 25,30,31.

Stochastic simulation algorithm
In the context of stochastic chemical kinetics, this refers to an exact discrete event simulation algorithm for generating time course trajectories of chemical reaction network models.

Monte Carlo error
The unavoidable error associated with estimating a population quantity from a finite number of stochastic samples from the population. It can often be reduced by averaging large numbers of samples.

Intrinsic noise
A crude categorization of stochasticity in biological systems that loosely corresponds to noise that cannot be controlled for.

Modelling intrinsic noise. Once stochastic models are created, they allow a range of investigations that are not possible using deterministic models. Modelling and experimental investigation of noise at the single-cell level in isogenic cell populations is currently the subject of many active research programmes. Stochastic modellers acknowledge the fact that molecules are discrete entities, and that reactions between molecules are stochastic events, which typically occur when molecules collide according to random processes. The state of the system at a given instant is therefore regarded to be a vector of counts of molecules, and remains constant until the next reaction event occurs. For the p53 example, this means that the state will be the actual number of molecules of p53, MDM2 and so on present in the cell; and this number will not change continuously, but will remain constant until changing abruptly each time a reaction involving those molecules occurs. For example, a p53–MDM2 binding event might occur when a p53 and MDM2 molecule collide at random in the cellular environment. The implications for the system state will be that the number of molecules of p53 and MDM2 will decrease by one, and the number of p53–MDM2 complexes will increase by one. Although it might seem reasonable to view the molecular dynamics of cellular processes as essentially deterministic, models concerned only with molecular counts do not explicitly consider the position and momentum of every single molecule, and so the timings of reaction events are essentially unpredictable.

Intrinsic noise in biochemical reactions has many components, including randomness of promoter binding and other DNA binding events, stochasticity in mRNA transcription and degradation processes, stochasticity of translation and protein degradation events, and randomness of other protein–protein and protein–metabolite interactions. Stochastic models allow investigation of the intrinsic variability of the cellular process of interest. For example, did the system evolve to suppress noisy gene expression? Such suppression could be achieved using a variety of techniques, including utilization of a carefully tuned signalling cascade[32]. Alternatively, has the system evolved to exploit noise? Several examples of noise exploitation are known for the Gram-positive bacterium *Bacillus subtilis* (BOX 3). Stochasticity in gene expression has been especially well studied in yeast (BOX 4), but it has also been observed and modelled in mammalian cells[14,33]. Experimental technology has developed to the extent that, in special cases, one can even observe stochasticity at the single-molecule level[34].

A common and well-studied example of noise exploitation is provided by bistability in a reaction network in conjunction with randomness of expression: this framework allows a single cell to select one of two phenotypic traits at random, with a probability that is specific to the network and to its associated initial conditions[35]. This allows organisms to express phenotypic heterogeneity even in uniform genetic and environmental conditions, and can have selective advantages. It is difficult to investigate such issues using deterministic models.

Other sources of heterogeneity. There are, of course, sources of variation in cellular systems other than intrinsic stochastic kinetic noise in biochemical reactions that should be incorporated into the models if they are to describe cell population behaviour effectively. One is randomness or uncertainty in the initial state of the biological system. For example, cells in a particular experiment might behave differently because they were different at the start of the experiment — perhaps having different RNA and protein levels — even if they are genetically identical.

One way to incorporate this uncertainty into the analysis is to construct simulations by first simulating initial conditions from a specified probability distribution, and then carrying out the stochastic simulation

Box 3 | **Exploitation of noisy gene expression by *Bacillus subtilis***

Noise in gene expression has particularly important consequences for small, single-celled prokaryotes. Like all organisms, they have evolved strategies for coping with noise. More interestingly, they have also evolved strategies for exploiting noise. Several interesting examples of these are exhibited by the Gram-positive bacterium *Bacillus subtilis*. Noise has a key role in the stochastic switching of the organism between vegetative and competent states[104,105]. Here, modelling shows that the existence of bistable fixed points (vegetative and competent) is not necessary, and that a model having one stable fixed point (vegetative) together with an excitable unstable one (competent) seems to be more consistent with experimental observations[104]. Noise associated with protein–DNA binding and unbinding is suggested as the key driver of the excitation dynamics.

Stochasticity in gene expression also seems to play an important part in the decision of whether to sporulate[106], thereby ensuring that only a small subset of cells in a population commit to spore formation. Intracellular noise also drives the transient heterogeneity of extracellular protease production[107].

In each case, it is optimal for the overall fitness of the isogenic cell population if only a small fraction of the individuals adopt the particular phenotypic trait, leading to a heterogeneous population containing many individuals that are well adapted to several possible changes in environment. *B. subtilis*, as many other single-celled organisms, uses noise to generate phenotypic heterogeneity in spite of uniform genetics and environmental conditions. In an uncertain world, this clearly improves the overall survival chances of the population. Stochasticity in gene expression is being used to generate inter-cell heterogeneity, which enables the population to cope with the stochasticity of the environment. Quantitative stochastic modelling is necessary to understand the relationship between the sources of noise and the distribution of population phenotypes.

Another phenotype adopted by only a small fraction of an isogenic population in good nutrient conditions is motility. Stochasticity is also key to enabling motile bacterial cells to navigate up nutritional gradients using chemotaxis. Each cell uses a 'tumble and swim' strategy — it switches randomly between a tumbling phase in which it randomly orients itself, and a swim phase, in which it swims in the direction it is currently oriented. By spending a longer, but still random, time swimming in directions where the nutritional concentrations increase, the cell effectively climbs the nutritional gradient. Intrinsic stochasticity in gene expression is the fundamental source of randomness in this strategy, and modelling can shed considerable light on the mechanisms[108].

algorithm to produce a trajectory that is dependent on that particular initial condition. Multiple independent realizations of the process constructed in this way will therefore incorporate both sources of uncertainty. Note that random initial conditions can be used in conjunction with continuous deterministic simulation models. That is, one can pick a random starting point, run a deterministic algorithm, and repeat the process to obtain a random ensemble of trajectories. However, this should not be regarded as an alternative to, or substitute for, carrying out stochastic simulation of noisy expression — the two sources of randomness are different, and can lead to qualitatively and quantitatively different behaviour. Taking the p53 model as an example, randomizing the initial conditions of a deterministic model of the network will clearly introduce some heterogeneity, but it will fail to capture the effect of the gradual loss of cell–cell synchronization as oscillations gradually and randomly drift out of phase. However, this effect is captured by the stochastic model even with fixed initial conditions.

Stochastic models are almost always required when a system is driven by random events. This is of particular relevance when modelling intracellular damage and repair processes, as damage often results from occasional low-frequency events such as ssDNA breaks caused by endogenous reactive oxygen species. The DNA repair processes are similarly stochastic, working perfectly most of the time but occasionally failing (with a particular probability). Such processes are especially relevant to the biochemical mechanisms of ageing (BOX 5).

There can also be uncertainty regarding kinetic rate constants, or it might be that the rate parameters of reactions are likely to vary randomly during the course of the simulation. The former can be dealt with in a similar manner to random initial conditions. The latter can arise if a reaction rate is modulated by a variable that is not explicitly being modelled; this situation should be handled by directly associating a time-varying stochastic process with the rate 'constant'. In practice, an independent noise process or a diffusion process is often used for this purpose. Either way, careful modification of the basic stochastic simulation algorithm is required[36]. Again, this approach can also be used in conjunction with continuous deterministic models for the system dynamics, but here the result is a fundamentally stochastic model. Indeed, such an approach was used by Geva-Zatorsky *et al.*[7] to capture the observed heterogeneity of p53 oscillations.

Fast stochastic models. The basic stochastic simulation algorithm becomes computationally infeasible for certain complex networks of practical interest — those with fast reactions or large numbers of certain biomolecules. Essentially, problems arise as soon as the model contains distinct processes operating on different timescales. For example, in the context of the p53 example, there can be thousands of p53–MDM2 binding and dissociation events for each p53 molecule synthesis event, and the stochasticity associated with the binding and dissociation can be small compared with that of synthesis, despite the fact that almost all of the computational effort is spent on simulating binding and dissociation events. Similar issues arise in continuous deterministic models. In this case, an alternative simulation algorithm is required. Although there are alternative exact simulation algorithms, such as the next reaction method[37], all exact algorithms become unusable in the context of challenging models.

Several approximate stochastic simulation algorithms have been proposed[38–45], and this is currently an active area of research[46–49]. Such algorithms generate time course trajectories from the model that have a probability distribution similar, but not identical, to that of the stochastic kinetic model. One obvious approach is to form an approximation of the Markov jump process model, and to simulate that. Although there are many possible ways to do this, I focus here on one particularly interesting approximation that is obtained as the diffusion approximation of the process, known in this context as the chemical Langevin equation (CLE)[50]. This approximation is based on finding a diffusion process (described by a stochastic differential equation, SDE) that closely matches the dynamics of the Markov jump process. It is usually straightforward to simulate realizations from the CLE using numerical integration schemes similar to those used for ordinary differential equations[51,52].

Diffusion process
A stochastic process continuous in both time and space and that can be described by a stochastic differential equation.

Next reaction method
An alternative exact simulation algorithm to the stochastic simulation algorithm, which in certain situations can be faster.

Diffusion approximation
A diffusion process that approximates a Markov jump process.

Chemical Langevin equation (CLE). A diffusion approximation to a stochastic chemical kinetic model.

Stochastic differential equation
(SDE). A mathematical equation involving both differential calculus and a stochastic process (typically Brownian motion). Simple cases can be 'solved' exactly, but typically solutions must be generated using a stochastic form of numerical integration.

Numerical integration
An algorithm (typically implemented on a computer) for generating approximate solutions to ordinary differential equations.

Multiscale model
A model that spans multiple scales in space and/or time. Such models generally require approximate algorithmic solutions, and are often computationally intensive.

Extrinsic noise
A crude categorization of stochasticity in biological systems that loosely corresponds to noise that can be controlled for.

Fluorescence-activated cell sorting
(FACS). An experimental technology that can be used to make quantitative measurements on a cell population with single-cell resolution. It is particularly useful for quantifying heterogeneity in cell populations.

Unlike Markov jump processes, SDEs have continuous trajectories. Although the state of the CLE is a vector of real numbers, the process retains all of the stochasticity associated with the discrete Markov jump process (BOX 1). Injecting additional sources of uncertainty that vary over time (such as time-varying reaction rate 'constants') is particularly convenient with a CLE model. In addition, the speed at which realizations of the CLE can be generated makes it particularly attractive for use in multiscale models. Of course, the CLE is only an approximation to the associated Markov jump process, and so the question of model accuracy naturally arises. The CLE is tolerably accurate, except in cases in which the system is being strongly driven by a molecule at very low copy number (zero, one or two molecules, for example). To keep things in perspective, it is helpful to bear in mind that the discrepancy between an 'approximate' and 'exact' model will typically be substantially less than the discrepancy between the 'exact' model and the real biological process.

The pragmatic approach adopted by many modellers is to begin by using an exact algorithm, and switch to an approximate algorithm only if computation time becomes prohibitive. As many simulation software packages incorporate both exact and approximate simulation algorithms, this is often a simple matter of selecting a different option. Unfortunately there is little theory that can provide reassurance about the accuracy of the approximate algorithms in challenging scenarios, but most perform reasonably well in practice[45]. In the context of the p53 model[14], it can take up to a couple of minutes of central processing unit (CPU) time on a fast computer to generate a single realization of 40 hours of simulation time using an exact algorithm such as the stochastic simulation algorithm. Various approximate algorithms can reduce this time significantly, depending on the accuracy required. Sampling from a CLE approximation can reduce simulation time by more than two orders of magnitude — reducing the CPU time to less than one second, albeit at the expense of an appreciable loss of accuracy.

Modelling across scales. Although much research effort is currently focused on stochasticity (both intrinsic and extrinsic noise[53]) at the single-cell level, the modelling framework extends readily to incorporate other important sources of heterogeneity at higher levels of biological organization. In the context of single-celled model organisms, this means heterogeneity in cell population behaviour that could be attributed to a variety of sources, including noise at the single-cell level, but also minor genetic and epigenetic variations, and variations in environment (such as crowding or intercellular signalling variation). For higher-level organisms, such as the mouse, additional sources of heterogeneity are even more important. Although the experimentalists' instinct to control for as many sources of potential heterogeneity as possible is well founded, we now understand that a degree of variation is unavoidable and must be incorporated into the models.

Through the development of integrated stochastic population models we can gain insight into the sources of heterogeneity in the system, and the extent to which noise in gene expression is propagated to observed heterogeneity at the population level[54,55]. Furthermore, by developing realistic models of cell population behaviour, we can realistically consider using data on cell populations, such as fluorescence-activated cell sorting (FACS) data, to calibrate the parameters of the single-cell models that drive the integrated cell population model. However, naively implemented multiscale stochastic models will be computationally prohibitive, so work will need to be done to speed up simulation models, either by making approximations or by exploiting high-performance computing facilities. Ultimately, there is a desire to develop this approach from simple cell population models to models of tissues, organs and, ultimately, multicellular organisms.

Top-down statistical modelling

In parallel to the developments in stochastic modelling of biological processes, increasing use is being made of statistical estimation procedures for fitting high-level descriptive (top-down) statistical models to experimental data. Statistical methods are widely used in genetics and bioinformatics, and provide a sophisticated framework for intelligent data analysis. In these areas, statistical models are often used to understand complex high-dimensional data sets. For example, they can help to identify candidate disease-causing genes on the basis of large genotyping data sets. Similarly, statistical techniques are used to identify genes that are differentially expressed, or perhaps cell cycle regulated, on the basis of microarray experiments.

Statistical models allow information to be extracted from data despite complex structures and noise processes in the data[56,57]. Bayesian methods[58,59] provide a particularly powerful framework for analysis[60–62]. The Bayesian approach, which will be described in greater detail in the subsequent sections, provides a fully probabilistic framework for describing models and prior knowledge about parameters, which leads naturally to sensible estimates of parameter values and associated levels of uncertainty.

In computational systems biology, a key goal of statistical modelling is to use high-throughput data to make inferences about the connectivity structure of the biological networks driving the data (for example, genetic

Box 4 | **Sources of stochasticity and heterogeneity in budding yeast**

The high degree of experimental tractability of *Saccharomyces cerevisiae*, and the availability of a variety of genome-wide libraries, makes it an ideal model for systematic study of noise in gene expression. There is compelling experimental evidence that noisy expression can often be detrimental to organismal fitness[109], and that there is selective pressure to minimize it[18]. It seems that genetic factors can modulate noise levels, and so stochasticity can be regarded as a complex genetic trait[19]. In general terms, noise seems to scale with protein abundance[110]. Yeast is also a good system for investigating sources of biological noise and heterogeneity. These include transcriptional noise, translational noise, minor genetic and epigenetic variations, micro-environmental variations and lack or loss of cell cycle synchrony. By using a genome-wide yeast GFP library in conjunction with high-throughput flow cytometry, there is now experimental evidence that the principal source of stochasticity in yeast protein expression is transcriptional noise associated with the production and degradation of mRNAs[111].

Box 5 | **Stochastic modelling of cellular ageing**

Many of the processes leading to cellular ageing are intrinsically random, and this makes ageing an especially appropriate target for stochastic modelling[112,113]. Although it is widely acknowledged that the ageing process is modulated by both genetic and environmental factors, the role of intrinsic chance is generally less well appreciated[3]. Nevertheless, in model organisms, the effect of chance seems to be considerable. For example, genetically identical *Caenorhabditis elegans* reared in uniform environmental conditions typically show as much as a twofold variation in lifespan across a population. Substantial variation in the lifespan of inbred laboratory mice strains is also observed. Although it is more difficult here to completely eliminate small genetic and environmental variations, it seems clear that intrinsic stochasticity again plays a significant part. The reason for this is that ageing is caused by low-frequency random events.

DNA damage occurs randomly, often caused by reactive oxygen species generated in the cell. Highly evolved repair pathways ensure the elimination of almost all damage, but occasional stochastic failures lead to the accumulation of random damage and the gradual loss of cellular function associated with ageing. Some basic mechanisms and principles can be studied in the context of model organisms such as budding yeast. For example, the accumulation of extrachromosomal ribosomal DNA circles has been implicated in yeast senescence[54], and the biochemical response to telomere uncapping has been modelled in some detail[114]. Models for mammalian cells, including the action of chaperones[115] and the p53–MDM2 system[14], give more direct insight into processes associated with human ageing. Modelling can help to unravel the complexities associated with many interacting damage and repair processes contributing to the ageing phenotype, and to help separate out the individual effects[116,117].

regulatory networks and signalling pathways). Although this problem is reasonably well defined, developing a satisfactory statistical framework for it is complex, largely owing to the combinatorial explosion of the number of possible connectivity structures as the number of potentially interacting partners increases. At a fundamental level, the problem remains largely unsolved. Nevertheless, there has been a great deal of useful work in this general direction.

Methods for Bayesian network inference have been applied to microarray data. For data that is not time course, the techniques range from simple learning algorithms through to fully probabilistic methods. Some techniques[63–67] require the conversion of continuous quantities to categorical ones (for example, high/low or up/down), which loses a lot of the information in the data. Conversely, methods that work with continuous data[68,69] are potentially more powerful, but rely on stronger modelling and distributional assumptions (that is, more aspects of the data need to be modelled). Although there are many limitations associated with these approaches, the techniques can provide useful insight into statistical associations between variables in the absence of time course data[67].

For time course data, such as some microarray experiments, there is a much greater potential for uncovering the causal influences driving the biological system dynamics, as it is possible to see how changes at one time point lead to changes in other properties at subsequent time points. Techniques used so far include methods for learning of dynamic Bayesian networks applied to discretized data[70,71] and methods for inferring continuous models[72]. Opgen-Rhein *et al.*[72] applied the methods to a time course microarray experiment to investigate the effect of the diurnal cycle on starch metabolism in *Arabidopsis thaliana*. The inferred network contained some highly connected nodes that are indicative of co-regulation of groups of genes. The continuous models rely on an assumption of linearity, but this does not seem to be a major limitation in practice. This is a promising new area for research, and I believe that there will be important developments in the next year or two, including fully Bayesian approaches to the problem, inspired by related work in other application areas, such as econometrics[73].

Although such statistical modelling techniques are useful for giving insight into possible network structure, the models are rather simple compared with the mechanistic stochastic models discussed in the earlier sections. In addition, the high-throughput data on which the models rely typically lack the resolution and precision to be useful for the quantitative estimation and calibration of detailed models of biochemical network dynamics.

Statistical inference for reaction networks

At the opposite end of the systems biology spectrum, there is considerable interest in using statistical methods to estimate parameters of detailed mechanistic (bottom-up) biological models using quantitative time course data on the system. For example, the mechanistic p53 model considered earlier contains many parameter values, such as initial conditions, reaction rates and binding constants, with values that are to some extent uncertain. The desire is to use time course experimental data to determine the values of these parameters that are most consistent with the data. These data are usually, although not necessarily[74], generated in a low-throughput manner, and typically involve measurements on only a small number of biochemical species (sometimes only one). It differs from typical high-throughput data as it has high time resolution, better calibration and lower experimental error. Ideally, these data will be at the level of a single cell[16,75], although this is not a fundamental requirement, especially for deterministic models[76]. For the p53 model, the single-cell fluorescence time series data on p53 and MDM2 levels (FIG. 1a) is ideal.

Bayesian inference for deterministic models. Researchers have been fitting deterministic models to time course data for decades, and many simple approaches are non-statistical[77]. The simplest approach involves defining a 'distance' between the model and the experimental data,

Bayesian methods
Fully probabilistic methods for describing models, parameters and data. So called because extensive use is made of Bayes theorem to compute the probability distribution of model parameters given the experimental data.

and then tuning uncertain parameters to minimize the distance measure. The simplest version of this is a non-linear least squares fitting approach. However, there are numerous problems with this approach that statistical methods can address. The statistical concept of likelihood provides the 'correct' way of understanding the discrepancy between the model and the experimental data[76]. In deterministic models, the likelihood approach coincides with the least squares approach precisely when all experimental measurement errors are assumed to be unbiased, independent and identically normally distributed. However, the likelihood concept provides a way of measuring distance in situations in which these strong assumptions do not hold.

Given that there is a (log–)likelihood function to optimize, there are other statistical issues that need to be considered even in the relatively simple case of deterministic models because of the nature of the likelihood surface. Typically, this is flat in the vicinity of the optimum, suggesting that there are a range of model parameters consistent with the data and casting doubt on the wisdom of focusing on a single point estimate. The likelihood surface is sometimes completely flat in a particular coordinate direction, indicating a lack of identifiability of the model. A related issue is that there are often 'ridges' in the likelihood surface corresponding to confounded parameters. Furthermore, the likelihood surface is often multimodal, with separated parts of parameter space providing adequate fits to the data. In addition, if there is some uncertainty regarding certain aspects of the model structure, then there is a statistical model selection problem that must somehow trade-off model fit against model complexity. All of the issues described here can be directly addressed using a Bayesian statistical approach using Markov chain Monte Carlo (MCMC) techniques[78–80].

Bayesian inference combines prior information regarding model parameters with information in the data (summarized by the likelihood function) to form a posterior distribution. This describes the uncertainty regarding model parameters that remains after having observed the experimental data[81]. Unfortunately this posterior probability distribution is, in general, analytically intractable for interesting problems. However, it is usually straightforward to construct MCMC algorithms[82–84] that explore this distribution and that can be used to compute any numerical summaries of interest. Thus, Bayesian methods provide much richer information about the relationship between the model parameters and the data than can be provided by a direct optimization approach to the parameter tuning problem. SloppyCell[78] and BioBayes[85] are freely available examples of general purpose software for parameter inference for deterministic models using Bayesian inference and MCMC.

The Bayesian approach also offers a clean solution to model selection. In the context of the p53 example, two competing models have been developed — one based on ATM signalling and one using p14ARF[14]. In the absence of concrete expert knowledge about which of these is more plausible for a given cell line, it would be useful to know which model is most consistent with the available data. Calculation of the Bayes factor using MCMC provides a quantitative answer to this question, and a computational solution for deterministic models is provided in REF. 80.

Inference for stochastic models. The benefits of statistical and Bayesian approaches to inference are especially apparent when fitting the parameters of single-cell stochastic kinetic models to time course experimental data. Stochastic models contain many rate constants, and some are less well studied than classical deterministic enzymatic rate constants, meaning that it is often difficult to find plausible values for them in the literature. However, inappropriate values for rate constants can often lead to poor behaviour of the model, both in terms of average behaviour and the characteristics of the system noise. When tuning the parameters of stochastic models, there is no obvious 'distance' function to optimize, owing to the fact that the likelihood function does not have a simple analytically tractable form. In this case, fairly sophisticated statistical analysis is required to make satisfactory progress with the parameter estimation problem. Non-Bayesian approaches to the problem must either try to approximate the likelihood[86], or use computationally intensive Monte Carlo methods to estimate it[87]. It is possible to develop exact Bayesian inference methods for this problem using MCMC[88], but the algorithms are computationally intensive and do not scale well to problems of realistic size and complexity. Although it is possible to speed up these algorithms somewhat by making some approximations[89], methods that are based around the Markov jump process and are associated with the stochastic kinetic model are likely to be problematic.

As previously discussed, replacing the exact stochastic model with an approximation (such as the CLE) can vastly reduce the computational problems associated with forwards simulation. The same technique can also be used for parameter inference. By replacing the Markov jump process with the CLE, the problem is changed from being that of estimating the parameters of a Markov jump process to one of estimating the parameters of a nonlinear multivariate diffusion process[90]. Although this is by no means a trivial problem, it is computationally amenable using sophisticated Bayesian inference techniques[91,92], and the resulting algorithms scale well to problems of realistic size and complexity[93,94]. It is interesting that the inference techniques based on the CLE can work well even in situations in which one would not expect the CLE to be a particularly good approximation to the Markov jump process, at least in terms of forward simulation (BOX 6). Although the theory for SDE parameter inference is now well developed, there are few examples of such techniques being applied to real biological data. It is to be anticipated that many applications will appear in the literature in the near future.

Calibration of multiscale stochastic models. The techniques discussed in the previous section are extremely powerful, and are of wide applicability to deterministic

Likelihood
The probability of the data given the statistical model and its parameters. In classical statistics it is often regarded as a function of the model parameters for given fixed experimental data.

Identifiability
The extent to which it is possible to accurately estimate model parameters given sufficient experimental data.

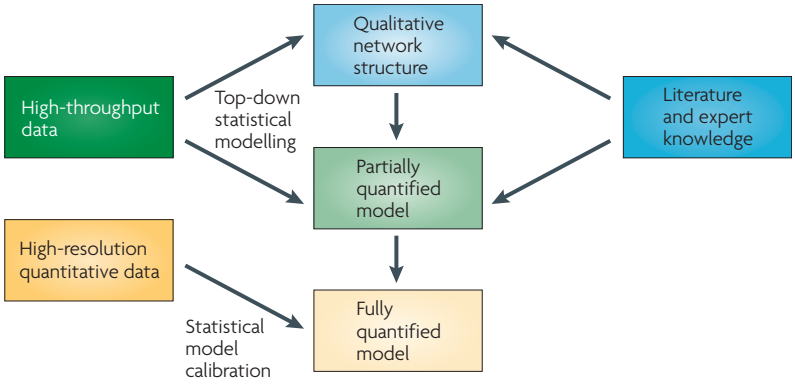
Confounded
Describes a problematic situation that arises when only a subset of a given set of model parameters is identifiable.

Model selection
The assessment of which model among a class of models has the most support on the basis of the available experimental data.

Markov chain Monte Carlo
(MCMC). A powerful class of algorithms that can be used to provide numerical solutions to most problems in Bayesian analysis. For complex problems they are notoriously computationally expensive, and many obscure techniques exist to increase the rate of convergence.

Posterior distribution
A probability distribution describing information about model parameters having taken into account all available information in the experimental data. From this it is possible to extract parameter estimates, together with associated levels of uncertainty.

Box 6 | **Statistical methods for biochemical network models**



Current approaches to using statistical methods to build quantitative dynamic models usually involve two or three stages. At the first stage, semi-quantitative high-throughput data is used to identify key biomolecules, and the nature of their interactions with other biomolecules (see figure). Dynamic statistical models can be used in conjunction with time course data to infer causal relationships. This information is used, together with existing literature, to form the qualitative structure of the reaction network. Data of this kind can also often be used to provide rough information regarding some model parameters, leading to a partially specified reaction network model. In the final stage, a quantitative reaction network model is calibrated against quantitative data to give a fully specified predictive dynamic model. Sophisticated statistical techniques are valuable at this stage too, especially if the model being calibrated is stochastic.

models and also to stochastic models using time course data at the single-cell level. However, there are many situations in which it is desirable to tune the parameters of a stochastic multiscale cell population model so that it better matches the data available at the population level, such as FACS measurements. In other application areas, the problem of tuning the parameters of computationally intensive simulation models is known as the model calibration problem[95]. MCMC algorithms typically take at least 10^6 times longer to run than it takes to forward simulate a single stochastic realization from the model. It is therefore unrealistic to expect to directly use such algorithms in the context of large, complex multiscale stochastic biological models.

In this case it is helpful to look at how the calibration of large and complex computer models is tackled in other disciplines, such as the modelling of oil reservoirs[96], weather forecasting and climate modelling[97]. In these scenarios, too, a direct fully probabilistic attack on the parameter inference problem is hopeless. However, sophisticated statistical procedures are still used in this context. In essence, output from the computer model that is obtained using a limited number of runs from different, and carefully chosen, parameter values is used to statistically estimate a fast emulator of the full model. This emulator can then be used as an approximate surrogate for the full model in any inferential procedure. In fact, this procedure can be a useful way of speeding up computations even in the context of a single cell, and was used to estimate the rate constants of the stochastic p53 model described earlier, using the available single-cell time course data[15].

There can be benefits associated with estimating the emulator and calibrating the model simultaneously[95]. Applications of these techniques to multiscale stochastic biological models are still in their infancy, but an example is given by Henderson *et al.*[98]. They used a single-cell model for the accumulation of mitochondrial DNA (mtDNA) deletions in a neuron of the substantia nigra region of the brain over time to generate a multiscale cell population model. An emulator is developed for this computationally intensive model, and the emulator is used in a Bayesian inference algorithm together with experimental data on mtDNA deletions in human brain tissue samples to infer key parameters of the single-cell model, such as the deletion rate and the lethal threshold for mtDNA deletions.

Fitting and exploiting stochastic emulators for multiscale model assessment and parameter inference is particularly challenging, as much of the necessary statistical theory has not yet been developed. However, the problem is of natural interest to statistical methodologists, and should therefore be an ideal opportunity for interdisciplinary collaboration.

Bridging the gap

It is of great interest to consider the possibility of developing an integrated framework for the simultaneous estimation of network structure and mechanistic model parameters using a combination of coarse-grained high-throughput data and fine-grained low-throughput time course data. For example, we might wish to extend or improve a mechanistic model of p53 oscillations using time course microarray data obtained from a population of cells. Such an approach is currently a daunting prospect, owing to the discrepancy between the high-level descriptive statistical models that are currently being used to analyse high-throughput data and the low-level mechanistic stochastic process models that are being used with fine-grained data, which represent the primary object of inferential interest. Again, the CLE offers a potential solution. The CLE connects directly with mechanistic models. However, it can be approximated by a tractable stochastic process known as an Ornstein–Uhlenbeck process[99]. This process can be time-discretized to give a model of the type sometimes used for inferring biological network connectivity[72]. There is therefore a natural sequence of directly comparable stochastic models, from bottom-up mechanistic to top-down descriptive, each of which can be linked to experimental data to differing degrees. This perhaps offers a glimpse of how fully integrated statistical models might be constructed in the future.

Conclusions

This article has looked at two related developments in computational systems biology. The first is the move towards the use of stochastic models for describing biological system dynamics, and the implications of this for more realistic and multiscale modelling. Realistic modelling of multiscale biological systems relies on the incorporation of the multiple sources of uncertainty, noise and heterogeneity that occur at different levels in the biological system. It is only by developing a fully

Emulator
A fast surrogate for a more complex, and hence slower, computational model. Emulators are often used in place of the original model in iterative algorithms that require many model evaluations.

Time-discretized
The conversion of a continuous time model to a discrete time model, formed by considering the states of the continuous time model only at given discrete times.

integrated model of this nature that experimental data at the whole-system level can be used effectively to estimate (that is, calibrate) model parameters and to assess the adequacy of the model. Therefore, completion of the iterative cycle of modelling and experimentation that is central to the systems biology approach actually requires integrated stochastic models of whole-system behaviour (see REFS 100,101 for a promising example). The development of such integrated multiscale models will require significant developments in stochastic simulation technology. The use of fast, approximate stochastic simulators will be necessary, as will the development of new techniques for simulating multiscale stochastic models. It is likely that statistically estimated stochastic emulators will be used for some model components in certain situations to reduce the computational demands of the algorithms. Techniques for running large stochastic simulation models on high-performance computing facilities will also require development.

The second area of this article concerned statistical estimation of network structure and model parameters. Here the challenges ahead are similarly formidable. Reliable simultaneous inference for network structure and kinetic parameters from a combination of high-throughput time course data and fine-grained time course data is a clear short-term goal. In the medium term, the development of techniques for effective calibration of large multiscale integrated stochastic models of complex biological systems is a key objective. In both cases sophisticated statistical methods will be required, and the problem structures make a Bayesian approach the obvious choice. It is therefore likely that we will see a Bayesian 'revolution' in computational systems biology, similar to that already experienced in genetics[102] and bioinformatics[103].

The scientific community must recognize the pivotal role of statistics and statisticians in systems biology research. No serious genetics laboratory or clinical trials unit would be considered complete without at least one expert statistical modeller. The contribution that a statistician can make to the success of a systems biology laboratory is every bit as great, but owing to the historical development of this new discipline, this fact has not been widely appreciated.

1. Kitano, H. Computational systems biology. *Nature* **420**, 206–210 (2002).
2. McAdams, H. H. and Arkin, A. It's a noisy business: genetic regulation at the nanomolecular scale. *Trends Genet.* **15**, 65–69 (1999).
3. Finch, C. E. & Kirkwood, T. B. L. *Chance Development and Aging* (Oxford Univ. Press 2000).
4. Maltzman, W. & Czyzyk, L. UV irradiation stimulates levels of p53 cellular tumor antigen in nontransformed mouse cells. *Mol. Cell. Biol.* **4**, 1689–1694 (1984).
5. Lev Bar-Or, R. *et al.* Generation of oscillations by the p53–mdm2 feedback loop: A theoretical and experimental study. *Proc. Natl Acad. Sci. USA* **97**, 11250–11255 (2000).
6. Lahav, G. *et al.* Dynamics of the p53–mdm2 feedback loop in individual cells. *Nature Genet.* **36**, 147–150 (2004).
7. Geva-Zatorsky, N. *et al.* Oscillations and variability in the p53 system. *Mol. Syst. Biol.* **2**, 2006.0033 (2006).
8. Haupt, Y., Maya, R., Kazaz, A. & Oren, M. Mdm2 promotes the rapid degradation of p53. *Nature* **387**, 296–299 (1997).
9. Clegg, H. V., Itahana, K. & Zhang, Y. Unlocking the mdm2–p53 loop: ubiquitin is the key. *Cell Cycle* **7**, 287–292 (2008).
10. Hucka, M. *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524–531 (2003).
11. Cornish-Bowden, A. *Fundamentals of Enzyme Kinetics* 3rd edn (Portland Press, 2004).
12. Ma, L. *et al.* A plausible model for the digital response of p53 to DNA damage. *Proc. Natl Acad. Sci. USA* **102**, 14266–14271 (2005).
13. Zhang, L. J., Yan, S. W. & Zhuo, Y. Z. A dynamical model of DNA-damage derived p53–mdm2 interaction. *Acta Physica Sinica* **56**, 2442–2447 (2007).
14. Proctor, C. J. & Gray, D. A. Explaining oscillations and variability in the p53–mdm2 system. *BMC Syst. Biol.* **2**, 75 (2008).
15. Henderson, D. A., Boys, R. J., Proctor, C. J. & Wilkinson, D. J. in *Handbook of Applied Bayesian Analysis* (eds O'Hagan, A. & West, M.) (Oxford Univ. Press) (in the press).
16. Bahcall, O. G. Single cell resolution in regulation of gene expression. *Mol. Syst. Biol.* **1**, 2005.0015 (2005).
17. Maheshri, N. & O'Shea, E. K. Living with noisy genes: how cells function reliably with inherent variability in gene expression. *Annu. Rev. Biophys. Biomol. Struct.* **36**, 413–434 (2007).
18. Lehner, B. Selection to minimise noise in living systems and its implications for the evolution of gene expression. *Mol. Syst. Biol.* **4**, 170 (2008).
19. Ansel, J. Cell-to-cell stochastic variation in gene expression is a complex genetic trait. *PLoS Genet.* **4**, e1000049 (2008).
20. Raser, J. M. & O'Shea, E. K. Noise in gene expression: origins, consequences, and control. *Science* **309**, 2010–2013 (2005).
21. Lopez-Maury, L., Marguerat, S. & Bahler, J. Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nature Rev. Genet.* **9**, 583–593 (2008).
22. Cox, D. R. & Miller, H. D. *The Theory of Stochastic Processes* (Chapman & Hall, London, 1977).
23. Gillespie, D. T. *Markov Processes: an Introduction for Physical Scientists* (Academic, New York, 1992).
24. Allen, L. J. S. *Stochastic Processes with Applications to Biology* (Pearson Prentice Hall, Upper Saddle River, 2003).
25. Wilkinson, D. J. *Stochastic Modelling for Systems Biology* (Chapman & Hall/CRC, Boca Raton, 2006).
26. Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361 (1977).
The original description of the stochastic simulation algorithm for discrete event simulation of biochemical reaction networks.
27. McAdams, H. H. & Arkin, A. Stochastic mechanisms in gene expression. *Proc. Natl Acad. Sci. USA* **94**, 814–819 (1997).
28. Zlokarnik, G. *et al.* Quantitation of transcription and clonal selection of single living cells with beta-lactamase as reporter. *Science* **279**, 84–88 (1998).
29. Renshaw, E. *Modelling Biological Populations in Space and Time* (Cambridge Univ. Press, 1991).
30. Li, H., Cao, Y., Petzold, L. R. & Gillespie, D. T. Algorithms and software for stochastic simulation of biochemical reacting systems. *Biotechnol. Prog.* **24**, 56–61 (2007).
31. Higham, D. J. Modeling and simulating chemical reactions. *SIAM Rev.* **50**, 347–368 (2008).
32. Paulsson, J., Berg, O. & Ehrenberg, M. Stochastic focusing: fluctuation-enhanced sensitivity of intracellular regulation. *Proc. Natl. Acad. Sci. USA* **97**, 7148–7153 (2000).
33. Dupont, G., Abou-Lovergne, A. & Combettes, L. Stochastic aspects of oscillatory Ca^{2+} dynamics in hepatocytes. *Biophys. J.* **95**, 2193–2202 (2008).
34. Cai, L., Friedman, N. & Xie, X. S. Stochastic protein expression in individual cells at the single molecule level. *Nature* **440**, 358–362 (2006).
35. Arkin, A., Ross, J. & McAdams, H. H. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected *Escherichia coli* cells. *Genetics* **149**, 1633–1648 (1998).
An important early example illustrating that stochastic kinetic models can describe important biological phenomena that cannot easily be understood using continuous deterministic models.
36. Shahrezaei, V., Ollivier, J. and Swain, P. Colored extrinsic fluctuations and stochastic gene expression. *Mol. Syst. Biol.* **4**, 196 (2008).
37. Gibson, M. A. & Bruck, J. Efficient exact stochastic simulation of chemical systems with many species and many channels. *J. Phys. Chem. A* **104**, 1876–1889 (2000).
38. Gillespie, D. T. Approximate accelerated stochastic simulation of chemically reacting systems. *J. Chem. Phys.* **115**, 1716–1732 (2001).
39. Gillespie, D. T. & Petzold, L. R. Improved leap-size selection for accelerated stochastic simulation. *J. Chem. Phys.* **119**, 8229–8234 (2003).
40. Kiehl, T. R., Mattheyses, R. M. & Simmons, M. K. Hybrid simulation of cellular behavior. *Bioinformatics* **20**, 316–322 (2004).
41. Alfonsi, A., Cances, E., Turinici, G., di Ventura, B. & Huisinga, W. Adaptive simulation of hybrid stochastic and deterministic models for biochemical systems. *ESAIM: Proc.* **14**, 1–13 (2005).
42. Puchalka, J. & Kierzek, A. M. Bridging the gap between stochastic and deterministic regimes in the kinetic simulations of the biochemical reaction networks. *Biophys. J.* **86**, 1357–1372 (2004).
43. Rao, C. V. & Arkin, A. P. Stochastic chemical kinetics and the quasi-steady-state assumption: application to the Gillespie algorithm. *J. Chem. Phys.* **118**, 4999–5010 (2003).
44. Haseltine, E. L. & Rawlings, J. B. Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. *J. Chem. Phys.***117**, 6959–6969 (2002).
45. Salis, H. & Kaznessis, Y. Accurate hybrid stochastic simulation of a system of coupled chemical or biochemical reactions. *J. Chem. Phys.* **122**, 054103 (2005).
46. Cao, Y., Gillespie, D. T. & Petzold, L. Multiscale stochastic simulation algorithm with stochastic partial equilibrium assumption for chemically reacting systems. *J. Comput. Phys.* **206**, 395–411 (2005).
47. Samant, A. & Vlachos, D. G. Overcoming stiffness in stochastic simulation stemming from partial equilibrium: a multiscale Monte Carlo algorithm. *J. Chem. Phys.* **123**, 144114 (2005).
48. Weinan, E., Liu, D. & Vanden-Eijnden, E. Nested stochastic simulation algorithm for chemical kinetic systems with disparate rates. *J. Chem. Phys.* **123**, 194107 (2005).
49. Weinan, E., Liu, D. & Vanden-Eijnden, E. Nested stochastic simulation algorithms for chemical kinetic systems with multiple time scales. *J. Comput. Phys.* **221**, 158–180 (2007).

50. Gillespie, D. T. The chemical Langevin equation. *J. Chem. Phys.* **113**, 297–306 (2000).
A well presented and accessible introduction to the chemical Langevin equation.
51. Cyganowski, S., Kloeden, P. & Ombach, J. *From Elementary Probability to Stochastic Differential Equations with MAPLE* (Springer, New York, 2002).
52. Kloeden, P. E. & Platen, E. *Numerical Solution of Stochastic Differential Equations* (Springer, New York, 1992).
53. Swain, P. S., Elowitz, M. B. & Siggia, E. D. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl Acad. Sci. USA* **99**, 12795–12800 (2002).
54. Gillespie, C. S. *et al.* A mathematical model of ageing in yeast. *J. Theor. Biol.* **44**, 493–516 (2004).
55. Tanase-Nicola, S. & ten Wolde, P. R. Regulatory control and the costs and benefits of biochemical noise. *PLoS Comput. Biol.* **4**, e1000125 (2008).
56. Speed, T. P. (ed.) *Statistical Analysis of Gene Expression Microarray Data* (Chapman & Hall/CRC, Boca Raton 2003).
57. Wit, E. & McClure, J. *Statistics for Microarrays: Design, Analysis and Inference* (Wiley, New York, 2004).
58. O'Hagan, A. & Forster, J. J. *Kendall's Advanced Theory of Statistics* Vol. 2B (Arnold, London, 2004).
59. Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. *Bayesian Data Analysis* 2nd edn (Chapman & Hall/CRC, Boca Raton, 2003).
60. Vanucci, M., Do, K.-A. & Muller, P. (eds) *Bayesian Inference for Gene Expression and Proteomics* (Cambridge Univ. Press, New York 2006).
61. Hein, A.-M. K., Richardson, S., Causton, H. C., Ambler, G. K. & Green, P. J. BGX: a fully Bayesian integrated approach to the analysis of Affymetrix Gene Chip data. *Biostatistics* **6**, 349–373 (2005).
62. Lewin, A., Richardson, S., Marshall, C., Glazier, A. & Aitman, T. Bayesian modelling of differential gene expression. *Biometrics* **62**, 10–18 (2006).
63. Friedman, N. Inferring cellular networks using probabilistic graphical models. *Science* **303**, 799–805 (2004).
64. Pournara, I. & Wernisch, L. Reconstruction of gene networks using Bayesian learning and manipulation experiments. *Bioinformatics* **20**, 2934–2942 (2004).
65. Beal, M. J., Falciani, F., Ghahramani, Z., Rangel, C. & Wild, D. L. A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics* **21**, 349–356, (2005).
66. Schafer, J. & Strimmer, K. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* **21**, 754–764 (2005).
67. Werhli, A. V., Grzegorczyk, M. & Husmeier, D. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. *Bioinformatics* **22**, 2523–2531 (2006).
68. Dobra, A. *et al.* Sparse graphical models for exploring gene expression data. *J. Multivar. Anal.* **90**, 196–212 (2004).
69. Jones, B. *et al.* Experiments in stochastic computation for high-dimensional graphical models. *Stat. Sci.* **20**, 388–400 (2005).
70. Husmeier, D. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics* **19**, 2271–2282 (2003).
71. Yu, J., Smith, V. A., Wang, P. P., Hartemink, A. J. & Jarvis, E. D. Advances to Bayesian network inference for generating causal networks from observational data. *Bioinformatics* **20**, 3594–3603 (2004).
72. Opgen-Rhein, R. & Strimmer, K. Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. *BMC Bioinformatics* **8** (Suppl. 2), S3 (2007).
The first paper to explore the use of sparse vector autoregressive models for inferring causal genetic regulatory relationships.
73. George, E., Sun, D. & Ni, S. Bayesian stochastic search for VAR model restrictions. *J. Econom.* **142**, 553–580 (2008).
74. Pepperkok, R. & Ellenberg, J. High-throughput fluorescence microscopy for systems biology. *Nature Rev. Mol. Cell Biol.* **7**, 690–696 (2006).
75. Shen, H. *et al.* Automated tracking of gene expression profiles in individual cells and cell compartments. *J. R. Soc. Interface* **3**, 787 (2006).
76. Jaqaman, K. & Danuser, G. Linking data to models: data regression. *Nature Rev. Mol. Cell Biol.* **7**, 813–819 (2006).
77. Moles, C. G., Mendes, P. & Banga, J. R. Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Res.* **13**, 2467–2474 (2003).
78. Brown, K. S. & Sethna, J. P. Statistical mechanical approaches to models with many poorly known parameters. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **68**, 021904 (2003).
An early example of applying MCMC methods for inferring parameters of continuous deterministic models.
79. Barenco, M. *et al.* Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biol.* **7**, R25 (2006).
80. Vyshemirsky, V. & Girolami, M. Bayesian ranking of biochemical system models. *Bioinformatics* **24**, 833 (2008).
Describes the use of MCMC for parameter inference and model selection using deterministic models.
81. Liebermeister, W. & Klipp, E. Biochemical networks with uncertain parameters. *IEE Syst. Biol.* **152**, 97–107 (2005).
82. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092 (1953).
83. Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109 (1970).
84. Gamerman, D. *Markov Chain Monte Carlo (Texts in Statistical Science)* (Chapman & Hall, New York, 1997).
85. Vyshemirsky, V. & Girolami, M. BioBayes: a software package for Bayesian inference in systems biology. *Bioinformatics* **24**, 1933–1934 (2008).
86. Reinker, S., Altman, R. M. & Timmer, J. Parameter estimation in stochastic biochemical reactions. *IEE Syst. Biol.* **153**, 168–178 (2006).
87. Tian, T., Xu, S., Gao, J. & Burrage, K. Simulated maximum likelihood method for estimating kinetic rates in gene expression. *Bioinformatics* **23**, 84–91 (2007).
88. Boys, R. J., Wilkinson, D. J. & Kirkwood, T. B. L. Bayesian inference for a discretely observed stochastic kinetic model. *Stat. Comput.* **18**, 125–135 (2008).
The first paper to demonstrate the possibility of conducting fully Bayesian inference for the parameters of stochastic kinetic models.
89. Rempala, G. A., Ramos, K. S. & Kalbfleisch, T. A stochastic model of gene transcription: an application to L1 retrotransposition events. *J. Theor. Biol.* **242**, 101–116 (2006).
90. Iacus, S. M. *Simulation and Inference for Stochastic Differential Equations — with R Examples* (Springer, New York, 2008).
91. Golightly, A. & Wilkinson, D. J. Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics* **61**, 781–788 (2005).
92. Heron, E. A., Finkenstadt, B. & Rand, D. A. Bayesian inference for dynamic transcriptional regulation; the Hes1 system as a case study. *Bioinformatics* **23**, 2596–2603 (2007).
93. Golightly, A. & Wilkinson, D. J. Bayesian sequential inference for stochastic kinetic biochemical network models. *J. Comput. Biol.* **13**, 838–851 (2006).
Describes using Bayesian inference for stochastic kinetic models using multiple, partial and noisy experimental data sets.
94. Golightly, A. & Wilkinson, D. J. Bayesian inference for nonlinear multivariate diffusion models observed with error. *Comput. Stat. Data Anal.* **52**, 1674–1693 (2008).
95. Kennedy, M. C. & O'Hagan, A. Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B* **63**, 425–464 (2001).
96. Goldstein, M. & Rougier, J. Bayes linear calibrated prediction for complex systems. *J. Am. Stat. Assoc.* **101**, 1132–114 (2006).
97. Challenor, P. G., Hankin, R. K. S. & Marsh, R. in *Avoiding Dangerous Climate Change* (Schellnhuber, H. J., Cramer, W., Nakicenovic, N., Wigley, T. & Yohe, G. eds) 53–63 (Cambridge Univ. Press, 2006).
98. Henderson, D. A., Boys, R. J., Krishnan, K. J., Lawless, C. & Wilkinson, D. J. Bayesian emulation and calibration of a stochastic computer model of mitochondrial DNA deletions in substantia nigra neurons. *J. Am. Stat. Assoc.* (in the press).
The first example of using inference for a single-cell model based on cell population data and a statistical emulator of a stochastic cell population model.
99. Uhlenbeck, G. E. & Ornstein, L. S. On the theory of Brownian motion. *Phys. Rev.* **36**, 823–841 (1930).
100. Orlando, D. *et al.* A probabilistic model for cell cycle distributions in synchrony experiments. *Cell Cycle* **6**, 478–488 (2007).
101. Orlando, D. *et al.* Global control of cell-cycle transcription by coupled CDK and network oscillators. *Nature* **453**, 944–947 (2008).
102. Beaumont, M. A. & Rannala, B. The Bayesian revolution in genetics. *Nature Rev. Genet.* **5**, 251–261 (2004).
103. Wilkinson, D. J. Bayesian methods in bioinformatics and computational systems biology. *Brief. Bioinformatics* **8**, 109–116 (2007).
104. Schultz, D., Jacob, E. B., Onuchic, J. N. & Wolynes, P. G. Molecular level stochastic model for competence cycles in *Bacillus subtilis*. *Proc. Natl Acad. Sci. USA* **104**, 17582–17587 (2007).
105. Smits, W. K. *et al.* Stripping *Bacillus*: ComK auto-stimulation is responsible for the bistable response in competence development. *Mol. Microbiol.* **56**, 604–614 (2005).
106. Veening, J.-W., Hamoen, L. W. & Kuipers, O. P. Phosphatases modulate the bistable sporulation gene expression pattern in *Bacillus subtilis*. *Mol. Microbiol.* **56**, 1481–1494 (2005).
107. Veening, J.-W. *et al.* Transient heterogeneity in extracellular protease production by *Bacillus subtilis*. *Mol. Syst. Biol.* **4**, 184 (2008).
108. Shimizu, T. S., Aksenov, S. V. & Bray, D. A spatially extended stochastic model of the bacterial chemotaxis signalling pathway. *J. Mol. Biol.* **329**, 291–309 (2003).
109. Fraser, H. B., Hirsh, A. E., Giaever, G., Kumm, J. & Eisen M. B. Noise minimization in eukaryotic gene expression. *PLoS Biol.* **2**, e137 (2004).
110. Bar-Even, A. *et al.* Noise in protein expression scales with natural protein abundance. *Nature Genet.* **38**, 636–643 (2006).
111. Newman, J. *et al.* Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* **441**, 840–846 (2006).
112. Kirkwood, T. B. L. *et al.* Towards an e-biology of ageing: integrating theory and data. *Nature Rev. Mol. Cell Biol.* **4**, 243–249 (2003).
113. Kirkwood, T. B. L. *et al.* in *Handbook of the Biology of Aging* 6th edn (eds Masoro, E. J. & Austad, S. N.) 334–357 (Academic, New York, 2005).
114. Proctor, C. J. *et al.* Modelling the checkpoint response to telomere uncapping in budding yeast. *J. R. Soc. Interface* **4**, 73–90 (2007).
115. Proctor, C. J. *et al.* Modelling the action of chaperones and their role in ageing. *Mech. Ageing Dev.* **126**, 119–131 (2005).
116. Kowald, A. & Kirkwood, T. B. Towards a network theory of ageing: a model combining the free radical theory and the protein error theory. *J. Theor. Biol.* **168**, 75–94 (1994).
117. de Sozou, P. & Kirkwood, T. B. L. A stochastic model of cell replicative senescence based on telomere shortening, oxidative stress, and somatic mutations in nuclear and mitochondrial DNA. *J. Theor. Biol.* **213**, 573 (2001).

Acknowledgements
The author would like to thank three anonymous referees for numerous suggestions that have helped to improve this article. This work was funded by the Biotechnology and Biological Sciences Research Council through grants BBF0235451, BBSB16550 and BBC0082001.

DATABASES
UniProtKB: http://www.uniprot.org
ATM | MDM2 | p14ARF | p53

FURTHER INFORMATION
Darren J. Wilkinson's homepage: http://www.staff.ncl.ac.uk/d.j.wilkinson
BioBayes: http://www.dcs.gla.ac.uk/BioBayes
Biology of ageing e-science integration and simulation system (BASIS): http://www.basis.ncl.ac.uk
BioModels database: http://www.ebi.ac.uk/biomodels
CaliBayes: http://www.calibayes.ncl.ac.uk
FERN — A Java Framework for Stochastic Simulation and Evaluation of Reaction Networks: http://www.bio.ifi.lmu.de/FERN
Systems Biology Markup Language (SBML): http://www.sbml.org
SloppyCell: http://sloppycell.sourceforge.net

ALL LINKS ARE ACTIVE IN THE ONLINE PDF