Vol 443|5 October 2006|doi:10.1038/nature05127

From *in vivo* to *in silico* biology and back

Barbara Di Ventura[1]*, Caroline Lemerle[1]*, Konstantinos Michalodimitrakis[1] & Luis Serrano[1]

The massive acquisition of data in molecular and cellular biology has led to the renaissance of an old topic: simulations of biological systems. Simulations, increasingly paired with experiments, are being successfully and routinely used by computational biologists to understand and predict the quantitative behaviour of complex systems, and to drive new experiments. Nevertheless, many experimentalists still consider simulations an esoteric discipline only for initiates. Suspicion towards simulations should dissipate as the limitations and advantages of their application are better appreciated, opening the door to their permanent adoption in everyday research.

> Intuition and concepts constitute, therefore, the elements of all our knowledge, so that neither concepts without an intuition in some way corresponding to them, nor intuition without concepts, can yield knowledge…Thoughts without content are empty, intuitions without concepts are blind…Only through their union can knowledge arise.[1]

In the past few years, the biological community has been exposed to a new buzzword: 'systems biology'. Irrespective of its exact definition, there are two attributes of systems biology on which most agree: an '-omics' aspect (involving the comprehensive collection of experimental data concerning a system), and the use of mathematical modelling to make testable predictions and gain insight about a biological system's behaviour. The intrusion of computational biology into 'wet' laboratories is producing a quiet revolution wherein simulation tools are used to complement experiments and accelerate the hypothesis generation and validation cycle of research[2–4]. Modelling a cellular process can highlight which experiments are likely to be the most informative in testing model hypotheses, and allow testing for the effect of drugs[5] or mutant phenotypes[6] on cellular processes—thus paving the way for individualized medicine[7].

Simulation of biological systems is a broad field, and the focus of this Review is mainly on networks operating at the cellular level[8,9]. Here, we expose the problems and limitations of models and simulations, and suggest when and how to use them. Our aim is to familiarize experimentalists with this exciting multidisciplinary endeavour so they might consider complementing their current tools with computational ones. As recent progress in the field has shown, computational biology can successfully assist experimentalists in unravelling the principles and operation of complex systems.

Modelling and simulation of biological systems

Biologists commonly use the term 'model' for verbal or graphical descriptions of a mechanism underlying a cellular process. Less often do they use it to refer to a set of equations expressing in a formal and exact manner the relations among variables that characterize the state of a biological system. The approach of biologists towards knowledge building has been mostly empirical (following 'intuition'), but experimental facts remain 'blind' without laws or principles derived from them. Conversely, theoretical approaches used by modellers have often failed to relate to real systems, such that theoretical concepts encapsulated in these studies are equally 'empty'. Instead, theory and experiments need to be viewed in close interplay.

Mathematical models are more rigorous and powerful than descriptive ones. In some cases, concepts derived from engineering—like the dampening effect of negative feedback on noise[10]—apply directly even to graphical models, but most often they provide nothing more than a mere indication of how a system might behave[11].

Bridging the gap between a large body of experimental data and a potentially useful mathematical model is not trivial. Knowledge about the system is essential and needs to be formalized for the chosen framework: for instance, by breaking reactions into mono- and bi-molecular ones if only elementary reactions are allowed. Ideally, all information relevant to a system (not only concentrations and rates of events, but also spatial distribution, diffusion parameters, excluded volumes, and so on) would be known to make a maximally accurate *in silico* replica of the system. Unfortunately, even for the best-studied systems, the mass of accumulated data still falls short of describing, even qualitatively, the variety of elementary processes that each molecular species engages in (post-translational modifications, degradation, complex formation, and so on); even less known are details of spatial information and the timing of events. Consequently, assumptions are necessary (for example, that all gene copies of a multi-copy plasmid are transcriptionally active, or that a certain molecule freely diffuses inside a cell or is always monomeric). On the other hand, it can be beneficial to exclude some known data to accommodate available computational power and to facilitate the analysis (even at the expense of accuracy). For example, irrelevant interactions of highly connected proteins could be omitted; details such as the cell-cycle regulation of a certain protein could be temporarily set aside; abundant species such as ATP or ribosomes might be represented as constant pools; or transcription and translation events might be lumped together.

With quantitative information often being both incomplete and of non-uniform quality, many modelling tools that are tailor-made to qualitative data are available (in addition to sometimes better known quantitative modelling approaches), some of which also give rise to quantitative predictions (for example, constraint-based modelling approaches). Given the limitations of any single modelling approach (mathematical tools applicable, data types allowed, and so on), it is often worthwhile trying a combination of them[12], and there are indeed a variety of diverse formalisms to choose from including some increasingly promising ones, such as Petri nets[13] and concurrency-theory-derived methods[14], inherited from computer science.

Qualitative modelling

The majority of current experimental techniques, including high-throughput ones, yield only qualitative or semiquantitative data. For

[1]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany.
*These authors contributed equally to this work.

simulations tools to be applied and useful in drawing non-obvious conclusions, the analysis of such data needs to allow, as a minimum, the formulation of logical statements describing, for instance, causal relationships between events involving model components. As an example, computer science algorithms used to perform code checks can assess the logical consistency of a set of statements: that is, check that no subset of statements is in contradiction with any other[15]. Automated tools such as these and others used in qualitative reasoning approaches become indispensable if logical inferences are to be made on very large sets of experimental observations. In qualitative modelling, kinetic processes[16] are simulated by tracking over discrete time the state of the system, defined in terms of a coarse range for each variable. The weak specification of such models conserves computer resources needed to explore the space of possible behaviours; moreover, it provides high-level predictions applying to a whole family of systems—for instance, the number of feedback loops or the ranges of variables supporting oscillations or switches. Although simulation of qualitative models can be fast, even a rough exploration of parameter space can become intractable as the size of the system increases, highlighting the need for increasing computer resources and methods to accelerate the parameters search.

In so far as biological systems have evolved tolerance to random fluctuations and perturbations, coarse ranges may suffice to predict correctly a system's behaviour[17]. For genes that are naturally found in only two states[18], the trade-off in accuracy may not even be high. On the other hand, simple models can, in some cases, predict behaviours that are far away from reality (Fig. 1 and below).

Constraint-based modelling

Many biological systems assume dynamic steady states, with the environment acting as source and sink for some molecules, whereas the concentration of the other molecules is balanced by the activities of the reactions (reaction fluxes) in the system. These states depend as much on specific properties of the components (for example, stoichiometries and reversibility of the reactions, enzymatic capacities) as they do on general constraints (for example, the law of mass action and laws of thermodynamics). Constraint-based modelling (Supplementary Fig. 1) incorporates this information in a framework that allows one to identify all sets of reaction fluxes that achieve steady state for a reaction network of interest. Built on precursor fields such as pathway analysis[19], this approach uses a rich variety of rigorous mathematical and computing tools (linear algebra, convex analysis and linear programming, to name but a few) to achieve results of high predictive value[20] along two lines of analysis: network-based pathway analysis (to extract systemic properties of the network such as pathway redundancy[21], missing reactions[22], and so on) and flux balance analysis[23] (to identify those system states in which so-called objective functions—for example, growth rate and ATP consumption—reach an optimal value). Adding constraints increases the accuracy of predictions by reducing the number of allowable states—for instance, by specifying conditions when reaction fluxes are zero because the corresponding enzymes are not expressed[24]. When a network's response to an environmental change is very fast compared with that change (that is, when the quasi-steady-state assumption holds) its time-evolution can be calculated recursively as a series of steady states (dynamic flux balance analysis; see Supplementary Fig. 1d). Having to consign reactions in a stoichiometry matrix, imposing the separation of substrates and products (for example, metabolites) from reaction effectors (for example, enzymes), means constraint-based modelling is rather inflexible; in particular, transcription regulatory gene networks and signal transduction cascades do not fit the formalism well. In fact, for them—as the quasi-steady state assumption is generally not valid and transient responses, inaccessible using this approach, are often the point of interest—other formalisms are often preferred.

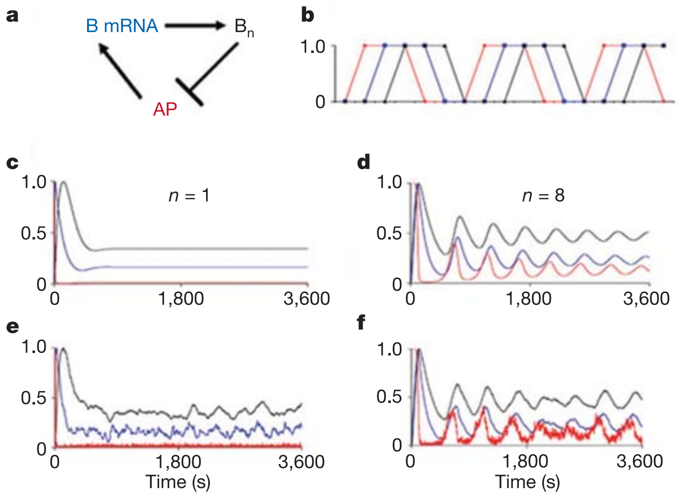

Figure 1 | Simulation of a simple network using different mathematical formalisms. **a**, Diagram of the negative feedback network used in the simulations. *n*, the number of B molecules in the active complex. **b–f**, Time courses of activator protein AP (red), *B* mRNA (blue) and B protein (black). The *y* axis represents the number of molecules, normalized for each species by the maximum value reached, except in **b**, in which it represents presence or absence of the molecules. Simulation of discrete time boolean model (**b**) with synchronous update. Deterministic (**c**, **d**) and stochastic[37] (**e**, **f**) simulations using specified parameters (see Supplementary Fig. 2), with B monomer (**c**, **e**) or octamer (**d**, **f**). Oscillations predicted by the boolean model are obtained in the deterministic/stochastic model only when B oligomerization is included.

Beyond the limited scope of application and the strong focus on steady states, a drawback of this method is the heavy computational load required to calculate optimal points, and its Achilles' heel is the choice of objective function[25]. Nevertheless, this technique has proved to be well-suited to analyse complex steady states and system responses to perturbations, whether genetic or environmental.

Quantitative modelling

Compared with qualitative models, quantitative ones have a natural appeal in that they offer greater detail in mimicking reality. Moreover, rich qualitative insights on the system are possible using theoretical tools such as bifurcation and stability analysis[26], which, for example, indicate the precise boundaries of parameter ranges to which steady states or sustained oscillations correspond, or reveal the stability of the solutions before actually solving the dynamical equations representing the system.

Quantitative models can be either deterministic or stochastic. The most popular formalism is the deterministic ordinary differential equations (ODEs) one which, when extended to model space, is referred to as partial differential equations (PDEs). Each equation in a set typically represents the rate of change of a species' continuous concentration as a sum or product of, more or less, empirical terms (typically law of mass action terms, Michaelis–Menten functions, power laws[27], and so on), accounting for the effect of biological events on such concentrations. By definition, the initial state of the system in a deterministic model uniquely sets all future states. As analytical solutions seldom exist, numerical solutions need then to be computed (once for each set of parameter values and initial conditions explored). A word of caution: although this step is simple in principle, wrong solutions can arise. For instance, the chosen steplength for the integration of the ODEs can be sufficiently large to cause divergence of the numerical solution from the correct one (numerical instability), making a minimum of experience with related issues a strong asset for the user. In general, ODEs are best suited to capturing the behaviour of systems where species are abundant and reaction events frequent (as is often the case for metabolic pathways, for example), because species concentrations are then acceptably approximated as varying continuously and predictably.

Molecular interactions are intrinsically random and cellular behaviour itself sometimes seems to reflect this randomness[28]. Indeed, occurrences of noise have been found to be exploited by cells—for instance, to survive a variety of environmental changes[29] or to increase sensitivity in signal transduction processes[30]. To model such stochastic systems, two main methods are used. The first comprises using stochastic differential equations (SDEs; derived from ODEs by adding noise terms to the equations), the solutions for which can be numerically obtained either by computing many trajectories (Monte Carlo methods) or approximating their probability distribution and then calculating statistical measures (such as mean and variance). Notably, with this method noise is imposed on the system and represented by mathematical terms chosen *a priori*, instead of arising from the underlying physical interactions. The second is a very successful and exact method introduced nearly 30 years ago[31,32], and recently enhanced to cope with different reaction timescales[33–35] or space[36–38]. With this approach, molecules are modelled individually and reaction events are calculated by their probability. The price to pay for having a more physically realistic model is the considerable increase in computational time and the need for specialized algorithms[36,37].

At the interface between purely deterministic and stochastic modelling techniques, novel hybrid approaches have appeared that speed up simulations considerably by partitioning reactions into fast reactions, assigned to a continuous framework, and slow reactions, simulated with a discrete stochastic algorithm[39,40].

Space in modelling

Until recently, the majority of simulations ignored the fact that biological processes take place in heterogeneous and highly structured environments. Even prokaryotes are now known to possess a cytoskeleton and control the movement and location of molecules, so regulating cellular processes in both space and time[41]. Indeed, spatial segregation underlies many cellular strategies; reactions are prevented by physically separating molecules, and molecular gradients[42] within or between cells are used in pattern formation.

Crucial as it may be for fundamental processes such as self-organization[43,44], morphogenesis[45], cell division[9] or calcium waves[46], spatial information is still largely absent from interaction databases. Recent technological advances are addressing this dearth of spatial data, and theoretical advances are improving computational methods, making it now possible to simulate spatio-temporal models of biological processes in coarse-grained or realistic geometries[47].

When to use different models: problems and solutions

Modelling could seem simple to an outsider: define your system, choose a modelling approach on the basis of what you know and want to know, download a simulation tool, input some parameters, run the simulation, and collect the results. However, as in the earlier days of protein design, what looks nice on a screen does not necessarily carry any biological meaning. There are many conceptual pitfalls for the modeller, which result in unrealistic predictions.

Mathematical-formalism-independent errors. Independent of the mathematical formalism used, many obvious errors can be introduced in a model. One set of errors arises from using parameter values measured in the often dilute, homogeneous environment of a test tube to model the crowded, apparently messy environment of cells. For example, the apparent binding constant of two interacting domains *in vivo* may differ from that obtained *in vitro* owing to crowding effects and/or spatial constraints[48]. Similarly, bulk estimates of the number of molecules per cell may hide local concentration variations that would affect reactions rates. Near the cell membrane, where the negative potential changes the electrostatic component, k_{on} values can be affected. Cell line, host and experimental conditions bring variability to data collected *in vivo* that needs to be considered.

Other subtle details can, if overlooked, lead to completely wrong predictions. For instance, assuming that the degradation rate of a protein is the same whether alone or in a complex is often unjustified, because complex formation can readily affect protein stability. In general, the common practice of assigning 'default' values to generic reactions (degradation rates, transcription activity, and so on) is acceptable only as a first approach and begs further refinement. The precise order in which a series of reactions occurs can be consequential, yet is frequently disregarded. For example (Fig. 2), if the formation of active dimers of a certain species is known to trigger downstream events, and, as an additional assumption, if such a dimer is only active when both monomers are activated (for example, by phosphorylation), one could formulate a general model considering all possible dimerization and monomer activation/deactivation reactions derived from these hypotheses. Not knowing the order of events, three extreme possibilities can be formulated: two of total dependence (only activated monomers can dimerize or dimerization must occur before activation can) and one of total independence between activation and dimerization reactions, leading to different steady-state concentrations of active dimer. Finally, when running a simple simulation over many cellular generations, surprises might be avoided by remembering to consider the effect of cell division on molecular pools.

Mathematical-formalism-dependent errors. Problems can arise from the mathematical formalism used to simulate a system. To illustrate the impact of modelling choices, simulations of a simple gene network with negative feedback (protein B forms multimers and sequesters the activator protein AP responsible for its transcription) were run using three formalisms with different degrees of graininess (Fig. 1): a simple boolean model, a quantitative deterministic model with ODEs, and a quantitative stochastic model[37]. With the discrete-time boolean model, the built-in delay produces oscillations (Fig. 1b).

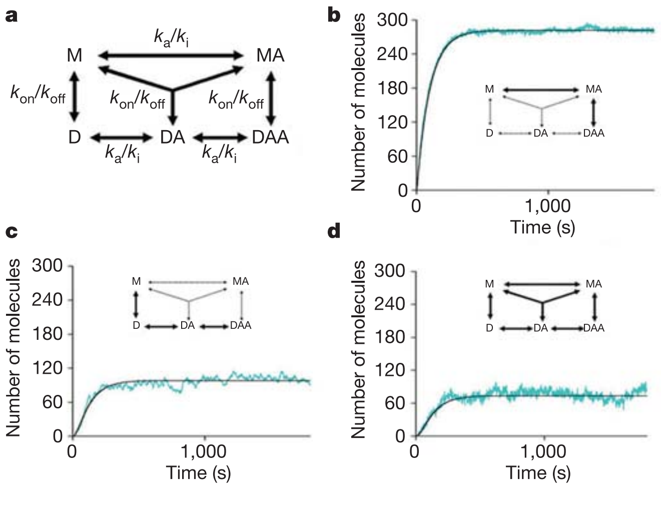

Figure 2 | Example of mathematical-formalism-independent pitfalls in modelling. If protein M is active as a post-translationally modified dimer DAA, the formation of DAA can be modelled differently, depending on assumptions made regarding the order of events and the nature of the active dimer. Here we assume that both monomers must be modified to form an active dimer. **a**, The scheme shows all reactions compatible with the experimental observations. M, monomer; MA, modified monomer; D, dimer; DA, dimer with one modified monomer; DAA, active dimer (both monomers modified). Each horizontal arrow corresponds to a reversible activation reaction and each vertical arrow corresponds to a reversible dimerization reaction. The two slanted arrows, together with the central vertical arrow, represent dimerization of a modified monomer with an unmodified one. **b**–**d**, Simulation runs for three different orders of events—respectively, 'activate then dimerize', 'dimerize then activate' and 'dimerize and activate together independently'—showing the deterministic (black) and stochastic (cyan) temporal evolution of DAA molecules (see Supplementary Fig. 3 for parameters used in the simulation and a mathematical derivation of the steady-state solution).

The other two models require additional events to be modelled explicitly (for example, degradation to balance production), and in contrast to what is observed with the coarser boolean model, oscillations did not occur unless multimerization was allowed (Fig. 1c–f).

Accounting for constraints of cellular space can significantly impact predictions, making spatial models more powerful compared with non-spatial ones. In Fig. 3, two different behaviours of a simple network are modelled, in which a phosphorylated transcription factor triggers the production, in one pole of the cell, of protein A that is involved in a positive feedback loop (resulting from mutual repression of species A and B). Starting from an initial state corresponding to a high concentration of B and no A, when the kinase and the phosphatase freely diffuse in the cell, the positive feedback acts as a switch causing the disappearance of B and the accumulation of A (Fig. 3d). When the kinase and the phosphatase are localized to opposite poles of the cell (Fig. 3b), however, a gradient of the phosphorylated transcription factor is formed (Fig. 3c), and the amount of A produced is insufficient to trigger the switch (Fig. 3e).

Because molecules are discrete species, continuous representations of molecular abundance are another source of artefacts. For example, contrary to what happens with discrete-valued models, steady states take infinite time to be reached with continuous concentrations, a discrepancy that disappears by focusing instead on how fast the steady state is approached (that is, by introducing the concepts of half-life, rise-time and others; Fig. 4a). Similarly, a probabilistic key is needed to interpret the non-integral values generated by the continuous-value deterministic models, as for discrete stochastic models. For low numbers, however, this interpretation is not error proof. If we consider a single, strictly autocatalytic species, (Fig. 4b), an ODE model would conclude (as would a related SDE model with a noise term that was only multiplicative) that there are two steady states: an unstable lower one (with zero molecules per cell) and a stable upper one. Simulation runs based on such models show that any non-zero initial state evolves asymptotically towards the upper steady state, whereas zero states are absorbing (that is, that once reached, they cannot be left). If important downstream events, such as cell differentiation or apoptosis, are triggered only by the absence of B molecules, a continuous model would lead us to wrongly conclude that these events never occur, whereas the more physically realistic, discrete stochastic model would reveal that, for each cell, it is only a matter of time before the triggering state is reached (Fig. 4b).

In summary, the modeller should never draw conclusions from simulation results without keeping in mind the limitations of a given approach to represent reality.

Consideration of the biological question. In the fortunate case in which both qualitative and quantitative information are available for our system and computational power is not limiting, the choice of formalism depends on which is best suited to capture the essential properties of the real system. Should time and variables in the model have continuous or discrete values? Does localization play an important part in any of the reactions? If so, which model geometry should be considered? Should fluctuations be allowed? Something that greatly helps in this decision process is having a clear biological question to answer and some ideas about the functional phenomenon being modelled. Thus, if the system includes gradients, the mathematical formalism used should handle space. If the system seems noisy (for example, not all cells respond in the same way to the same stimulus), then a stochastic approach might clarify this point. For metabolic pathways especially, constraint-based modelling provides a wealth of avenues to explore. Although these examples may induce wariness in the novice modeller when trying to decide on an approach, such considerations will become second nature as more practice in the 'systems biology' mentality is gained.

Successes in modelling

A model should allow the behaviour of a system to be predicted under various conditions, including some not yet experimentally tested and therefore not exploited in the model-building process. This most often requires an iterative process of model building and experimental investigation. Although most current models are at an early stage of this process, the number of cases where the symbiotic interaction of experiment and model has led to successful predictions is steadily growing. The first examples came in the field of metabolic engineering, where modelling led to the identification of bottlenecks in production[49]. More recently, as constraint-based modelling approaches are applied to reconstructed networks of genome scale, significant progress is being made in understanding and predicting the phenotypic potential of microorganisms[50,51].

Models have been most successful for systems of simpler organisms, like *Escherichia coli* and *Saccharomyces cerevisiae*. Two interesting examples are those of sphingolipid metabolism[52] and the osmotic shock response[53] in *S. cerevisiae*, with the latter being a good example of

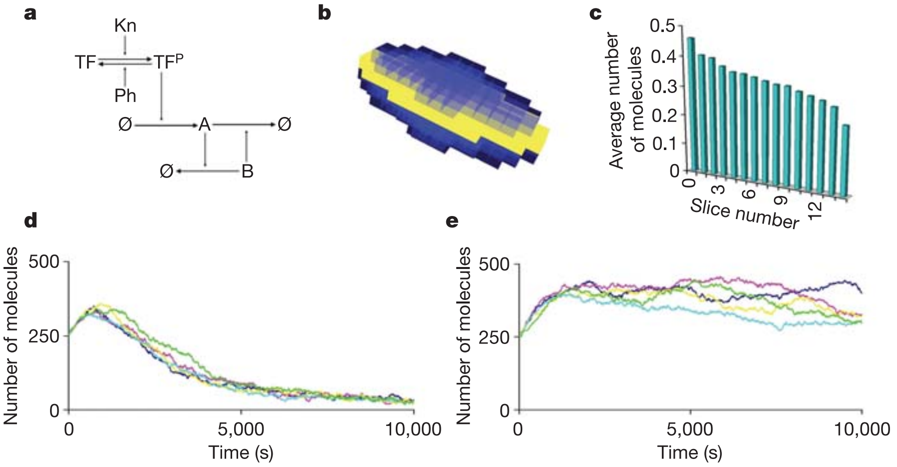

Figure 3 | Effect of localization of species on cellular processes. **a**, Diagram of a simple network in which a phosphorylated transcription factor TF^P triggers the synthesis of protein A that is involved in a positive feedback loop with protein B (resulting from mutual repression). Ø → indicates protein production, whereas → Ø indicates protein degradation. The behaviour will depend on the spatial constraints imposed on the kinase Kn and the phosphatase Ph. When they localize to opposite poles of the cell, switching behaviour does not take place. Production of protein A in both simulations[37] takes place in the pole of the cell where the phosphatase is localized. **b**, The model geometry (lattice unit is $1\,\mu$m). Yellow indicates the longitudinal cell slice along which the gradient is observed. **c**, Gradient of TF^P obtained with localization of species. **d**, **e**, Time course of protein B without (**d**) and with (**e**) localization of species (see Supplementary Fig. 4 for parameters used in the simulation).

overcoming data limitation by dividing the system into functional modules. One of the most beautiful paradigms of simulation–experiment interplay is in *E. coli* chemotaxis[54,55], but also significant discoveries have been made thanks to mathematical modelling in mammalian epidermal growth factor receptor (EGFR) and extracellular-signal-regulated kinase (ERK) signalling pathways[56,57].

As these recent examples show, the interplay between experiments and modelling has led to increasingly accurate models that not only fit the experimental data, but also make correct predictions and suggest new testable hypotheses.

Gaining understanding by mimicking nature

Natural systems can be discouragingly complex when tackled from a reverse-engineering approach. An alternative approach that makes extensive use of modelling and simulation tools has recently shown great promise: that of designing and building synthetic networks. Similar to the field of protein folding, where proteins are being successfully redesigned without a full understanding of how the proteins fold, biological circuits are being designed to engineer new properties into living organisms, despite the limitations in our understanding of the cell processes.

Designs found in nature can be tested in isolation for a given function (Fig. 5). Well-characterized components help lower the barriers to modelling. The use of control elements (such as temperature for a temperature-sensitive protein, or an exogenous small molecule affecting a reaction) helps model validation by allowing a multitude of conditions to be tested. Also, the possibility to connect synthetic networks (for example, connecting a switch with a rheostat to make an oscillator) paves the way for investigating principles of modularity.

Synthetic networks are often patchworks of previously used parts, not only because a single point mutation may alter the *in vivo* activity of the network, but also because we cannot currently predict how redesigned molecules such as synthetic promoters will behave. Indeed, the better characterized the pieces are, the more confident one can be in the corresponding models (although spurious interactions can always arise owing to the complexity of cellular environments causing discrepancies between models and experiments). In fact, a central repository for well-characterized parts and modules and related information already exists (http://parts.mit.edu/registry/index.php/Main_Page), and is commendably helping to consolidate and expand knowledge in this field.

Designing a synthetic biological network involves a conscious choice of inputs and outputs—the handles and readouts of the system. As an example of a handle, compounds can be added that modulate the behaviour of specific components (for example, the Tet repressor can be inactivated by anhydrotetracycline), allowing the exploration of many configurations, including unnatural ones;

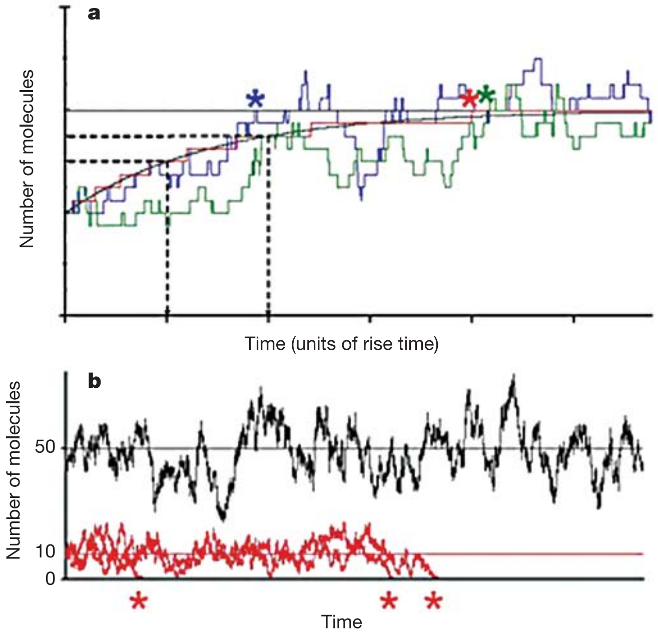

Figure 4 | Example of putative model-dependent pitfalls in modelling: continuous versus discrete concentrations. **a**, Simulation of a simple system in which a species is produced at a constant rate (16×10^{-3} molecules s^{-1}), and degraded with a first order decay rate of 1×10^{-3} s^{-1} giving an asymptotic steady-state of 16 molecules per cell (plotted as a straight horizontal line). The initial state for all simulations is 8 molecules per cell. Two discrete stochastic runs are shown in blue and green. The corresponding ODE solution is shown in black, whereas the discrete approximation of it is shown in red (rounding each value of the continuous run to the closest integer). A star indicates, for each discrete run, when the steady-state value has been reached. Each broken line illustrates a geometric construction of the rise-time, as the time taken by the system to go, first from 8 to 4 molecules per cell away from steady state, then from 4 to 2. Each unit of rise-time is separated by tick-marks on the *x*-axis, which are identically spaced (as a property of exponential curves). The stochastic simulation shows that it is possible to reach steady-state values earlier or later than in the deterministic analysis. **b**, Superimposed stochastic (wavy lines) and deterministic (continuous straight lines) simulation runs of a system composed of a single autocatalytic species, with parameter values (see Supplementary Fig. 5) such that the upper steady-state solution is either high (black curves) or low (red curves) (respectively, 50 and 10 molecules per cell volume of 1 μm^3). A star indicates when a stochastic run has reached the zero absorbing state. In the lower case (red curves), the deterministic solution clearly differs from the average stochastic run, because all runs reach zero and stay there.

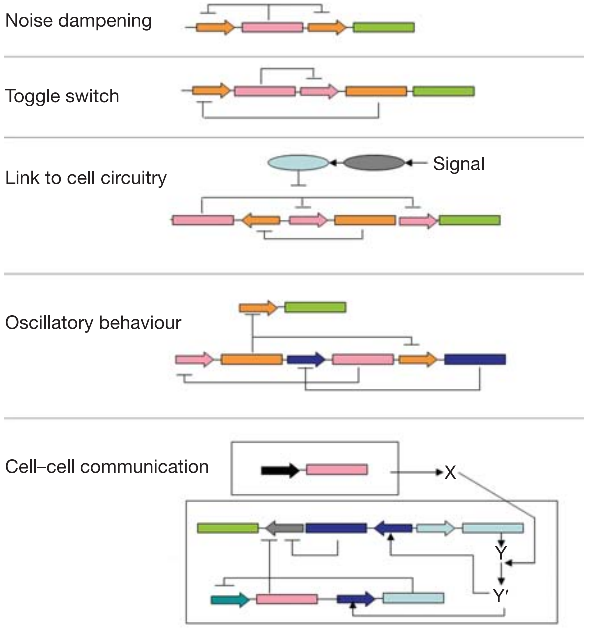

Figure 5 | Example of synthetic gene networks built with defined behaviour. Promoters are represented by arrows and genes by elongated boxes. Products of genes that are activated by another product are represented by X or Y. Cells are represented by large transparent boxes. Natural cell circuits are represented by ellipses. Reporter products are normally shown in green. When more than one bibliographic reference is given, only the circuitry used in the first cited reference is shown. Shown are: a noise-dampening negative feedback loop[10]; switch behaviour (toggle switch mutual repression[65–67], positive feedback loop[68,69] and logic gates[11]); interfacing natural and synthetic networks[70]; oscillatory behaviour (repressilator[71], metabolic oscillator[72] and oscillator resistant to fluctuations[73]); and cell–cell communication (regulation of protein expression[74], biofilm formation[70], population dynamics[75], selective killing of cells[75], pattern formation[76] and detection of gene transfer[77]).

although the list of such known regulator/molecular target pairs is still limited, it is being expanded by dedicated efforts (for example, drug design). At the other end, the experimental readout is expected to give a static or dynamic picture of the state of the system, for a cell or a population, usually measuring the concentrations of key components. Appropriate detection methods are those that offer a good compromise between non-invasiveness, sensitivity, versatility, detection speed, ease of measurement, and so on, with champions of the trade such as green fluorescent protein, promising new candidates such as fluorescein arsenical hairpin binder (FlAsH)[58], and also a wealth of antibody and radioisotope-based methods to choose from. Finally, as most detection methods are indirect, carefully chosen internal standards are needed for comparing experimental data to model predictions (for example, fluorescence counts and number of molecules per cell).

For years, models of gene networks have been used to investigate general issues of relevance to molecular biology, but were awaiting proper experimental validation. The field of synthetic biology has clearly started addressing these issues with circuits of various designs[59] (Fig. 5) guided by their own collection of models, and many experimental results are corroborating simulation predictions.

From simulating to understanding: design principles

Contemporary experimental techniques offer insight about biological functions at the molecular level. Given the complexity of natural systems, however, knowledge of all interactions happening at this level does not generally provide the modeller with an intuitive and coherent comprehension about the process of interest. For instance, although we can understand how a signal is propagated in a linear cascade from cell membrane to nucleus, we find it difficult to make sense of a highly interconnected gene network, with the added complication that only part of the network may be active at any one time. One strategy is to decompose the network into more manageable modules, a task that is not trivial and somewhat arbitrary[60] because it requires grouping parts on the basis of the purported biological function achieved by the module within the network. Groupings are primarily suggested by topology, but as mentioned above, cases abound in which topology alone is insufficient to elicit function. Even close-to-ideal, single-input, single-output modules with a clear function, such as some signal transduction cascades seem to be, can require further specifications to predict important details of their operation, such as the topology-inferred property of ultrasensitivity, which was recently shown to depend on a set of conditions usually not met in real cascades[61].

The search for and characterization of modular units within networks is an active field of systems biology where the observation of recurring designs linked to specific properties is providing another level of description of biological systems. A behaviour predicted for an isolated system, such as a small network motif (recurring topological pattern found over-represented in biological networks[62]), is not straightforward to extrapolate to systems in their natural environment, given how intricately embedded they can be; consequently, despite the established modularity of networks, successful extrapolations[63] almost come as a surprise. These higher-level descriptions of how biological network modules behave draw heavily on analogies with the way human-engineered systems work. Engineering concepts about network components (for example, switches, amplifiers and control elements) and properties (for example, robustness and modularity) have provided invaluable rational descriptions of module behaviour by abstracting the details of molecular interactions.

Pushing the engineering analogy even further, systems biology studies have the aim of uncovering the ‘organizational principles’, or ‘design principles’, of biological systems. Although there are obvious radical differences between human-engineered and natural systems (for instance, engineers lay down a wish-list of intended properties for their system during the design phase), natural systems do have solutions similar to human-engineered ones in terms of certain emergent properties (for example, modularity and noise attenuation), details of design (for example, feedback loops) and behaviour (for example, oscillations), as if conforming to a strict set of constraints that biologists are on a quest to discover (not to be confused with any universal law of nature, embedded, as if by magic, in each and every life form). Actually, beyond their natural appeal, the use of systems-theoretical concepts is perhaps our only chance to logically formulate the way a complex biological process operates in a concise, synthetic, human-understandable manner.

Although much remains to be learned in this field, simulation predictions of natural or engineered biological networks are helping us to identify the logical links between system design and system behaviour. This includes understanding the rationale behind the preferential use in nature of certain molecular blueprints to elicit certain elementary functions, such as the use of negative feedback loops to achieve homeostasis[13], sign-sensitive delay properties of feed-forward loops[63], or more elaborate designs such as robust entrainment and temperature compensation of circadian clocks (achieved through densely connected loops[64]), and the remarkable property of cellular ‘adaptation’, so simply formulated, and yet still so mysterious[56].

Perspectives

By discovering design principles, identifying biological modules, and quantitatively understanding how they operate through experiments and simulations, we hope to elucidate biological function as well as predict the effect of internal perturbations (for example, genetic mutations) or external perturbations (for example, drugs) so that disease treatments are more precise and effective. Similarly, understanding biological modules and being able to engineer new ones will pave the way for re-engineering of organisms and cells for numerous applications (including medical, agricultural and ecological situations). Thus, although a global and perfect understanding of a living system is not expected in the near future, the combination of modelling and experimentation offers the possibility of making inroads towards that goal, as well as developing new exciting, useful applications.

1. Kant, I. *Critique of Pure Reason*, University of Virginia Library, Electronic Text Center, Topic I, Part II, 45 ⟨http://etext.lib.virginia.edu/toc/modeng/public/KanPure.html⟩.
2. Locke, J. C. W. Extension of a genetic network model by iterative experimentation and mathematical analysis. *Mol. Syst. Biol.* doi:10.1038/msb4100018 (28 June 2005).
3. Albert, M. A. *et al.* Experimental and *in silico* analyses of glycolytic flux control in bloodstream form *Trypanosoma brucei*. *J. Biol. Chem.* **280**, 28306–28315 (2005).
4. Lee, E., Salic, A., Kruger, R., Heinrich, R. & Kirschner, M. W. The roles of APC and Axin derived from experimental and theoretical analysis of the Wnt pathway. *PLoS Biol.* doi:10.1371/journal.pbio.0000010 (13 October 2003).
5. di Bernardo, D. *et al.* Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nature Biotechnol.* **23**, 377–383 (2005).
6. Segre, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl Acad. Sci. USA* **99**, 15112–15117 (2002).
7. Chang, P. L. Clinical bioinformatics. *Chang Gung Med. J.* **28**, 201–211 (2005).
8. Kerckhoffs, R. C. *et al.* Electromechanics of the paced left ventricle simulated by a straightforward mathematical model: comparison with experiments. *Am. J. Physiol. Heart Circ. Physiol.* **5**, H1889–H1897 (2005).
9. Dens, E. J., Bernaerts, K., Standaert, A. R. & Van Impe, J. F. Cell division theory and individual-based modeling of microbial lag: part I. The theory of cell division. *Int. J. Food Microbiol.* **101**, 303–318 (2005).
10. Becskei, A. & Serrano, L. Engineering stability in gene networks by autoregulation. *Nature* **405**, 590–593 (2000).
11. Guet, C. C., Elowitz, M. B., Hsing, W. & Leibler, S. Combinatorial synthesis of genetic networks. *Science* **296**, 1466–1470 (2002).
12. Shmulevich, I., Kauffman, S. A. & Aldana, M. Eukaryotic cells are dynamically ordered or critical but not chaotic. *Proc. Natl Acad. Sci. USA* **102**, 13439–13444 (2005).
13. Hardy, S. & Robillard, P. N. Modeling and simulation of molecular biology systems using Petri nets: modeling goals of various approaches. *J. Bioinform. Comput. Biol.* **2**, 595–613 (2004).
14. Errampalli, D. D., Priami, C. & Quaglia, P. A formal language for computational systems biology. *OMICS* **8**, 370–380 (2004).

15. Batt, G. *et al.* Validation of qualitative models of genetic regulatory networks by model checking: analysis of the nutritional stress response in *Escherichia coli*. *Bioinformatics* **21** (Suppl 1), i19–i28 (2005).
16. Kuipers, B. in *Readings in Qualitative Reasoning about Physical Systems* (ed. deKleer, D. S. W. J.) 257–274 (Morgan Kaufmann, San Francisco, 1989).
17. Csete, M. E. & Doyle, J. C. Reverse engineering of biological complexity. *Science* **295,** 1664–1669 (2002).
18. Louis, M. & Becskei, A. Binary and graded responses in gene networks. *Sci. STKE* **143,** PE33 (2002).
19. Clarke, B. L. Complete set of steady states for the general stoichiometric dynamical system. *J. Chem. Phys.* **75,** 4970–4979 (1981).
20. Edwards, J. S., Ibarra, R. U. & Palsson, B. O. *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nature Biotechnol.* **19,** 125–130 (2001).
21. Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S. & Gilles, E. D. Metabolic network structure determines key aspects of functionality and regulation. *Nature* **420,** 190–193 (2002).
22. Reed, J. L., Vo, T. D., Schilling, C. H. & Palsson, B. O. An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol.* **4,** R54 (2003).
23. Fell, D. A. & Small, J. R. Fat synthesis in adipose tissue. An examination of stoichiometric constraints. *Biochem. J.* **238,** 781–786 (1986).
24. Covert, M. W., Schilling, C. H. & Palsson, B. Regulation of gene expression in flux balance models of metabolism. *J. Theor. Biol.* **213,** 73–88 (2001).
25. Burgard, A. P. & Maranas, C. D. Optimization-based framework for inferring and testing hypothesized metabolic objective functions. *Biotechnol. Bioeng.* **82,** 670–677 (2003).
26. Fall, C. P., Marland, E. S., Wagner, J. M. & Tyson, J. J. (eds) *Computational Cell Biology*. 1st edn. (Springer, 2002).
27. Savageau, M. A. Biochemical systems theory: operational differences among variant representations and their significance. *J. Theor. Biol.* **151,** 509–530 (1991).
28. Liu, Q. & Jia, Y. Fluctuations-induced switch in the gene transcriptional regulatory system. *Phys. Rev. E* **70,** 041907 (2004).
29. Thattai, M. & van Oudenaarden, A. Stochastic gene expression in fluctuating environments. *Genetics* **167,** 523–530 (2004).
30. Hanggi, P. Stochastic resonance in biology. How noise can enhance detection of weak signals and help improve biological information processing. *ChemPhysChem* **3,** 285–290 (2002).
31. Gillespie, D. T. General method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* **22,** 403–434 (1976).
32. Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81,** 2340–2361 (1977).
33. 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).
34. Rathinam, M., Petzold, L. R., Cao, Y. & Gillespie, D. T. Stiffness in stochastic chemically reacting systems: the implicit tau-leaping method. *J. Chem. Phys.* **119,** 12784–12794 (2003).
35. 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).
36. Stundzia, A. B. & Lumsden, C. J. Stochastic simulation of coupled reaction-diffusion processes. *J. Comput. Phys.* **127,** 196–207 (1996).
37. Ander, M. *et al.* SmartCell, a framework to simulate cellular processes that combines stochastic approximation with diffusion and localisation: analysis of simple networks. *Systems Biol.* **1,** 129–138 (2004).
38. Bezrukov, S. M., Frauenfelder, H. & Moss, F. (eds) *Fluctuations and Noise in Biological, Biophysical, and Biomedical Systems* (Proc. SPIE, Vol. 5110, 2003).
39. Salis, H. & Kaznessis, Y. Accurate hybrid stochastic simulation of a system of coupled chemical or biochemical reactions. *J. Chem. Phys.* **122,** 054103 (2005).
40. 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 doi:10.1051/proc:2005001 (2005).
41. Gitai, Z. The new bacterial cell biology: moving parts and subcellular architecture. *Cell* **120,** 577–586 (2005).
42. Gorlich, D., Seewald, M. J. & Ribbeck, K. Characterization of Ran-driven cargo transport and the RanGTPase system by kinetic measurements and computer simulation. *EMBO J.* **22,** 1088–1100 (2003).
43. Nedelec, F., Surrey, T. & Karsenti, E. Self-organisation and forces in the microtubule cytoskeleton. *Curr. Opin. Cell Biol.* **15,** 118–124 (2003).
44. Sawai, S., Thomason, P. A. & Cox, E. C. An autoregulatory circuit for long-range self-organization in *Dictyostelium* cell populations. *Nature* **433,** 323–326 (2005).
45. Collier, J. R., Monk, N. A., Maini, P. K. & Lewis, J. H. Pattern formation by lateral inhibition with feedback: a mathematical model of Delta–Notch intercellular signalling. *J. Theor. Biol.* **183,** 429–446 (1996).
46. Wu, D., Jia, Y., Yang, L., Liu, Q. & Zhan, X. Phase synchronization and coherence resonance of stochastic calcium oscillations in coupled hepatocytes. *Biophys. Chem.* **115,** 37–47 (2005).
47. Lemerle, C., Di Ventura, B. & Serrano, L. Space as the final frontier in stochastic simulations of biological systems. *FEBS Lett.* **579,** 1789–1794 (2005).
48. Ellis, R. J. Macromolecular crowding: obvious but underappreciated. *Trends Biochem. Sci.* **26,** 597–604 (2001).
49. Yarmush, M. L. & Banta, S. Metabolic engineering: advances in modeling and intervention in health and disease. *Annu. Rev. Biomed. Eng.* **5,** 349–381 (2003).
50. Price, N. D., Reed, J. L. & Palsson, B. O. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Rev. Microbiol.* **2,** 886–897 (2004).
51. Fong, S. S. & Palsson, B. O. Metabolic gene-deletion strains of *Escherichia coli* evolve to computationally predicted growth phenotypes. *Nature Genet.* **36,** 1056–1058 (2004).
52. Alvarez-Vasquez, F. *et al.* Simulation and validation of modelled sphingolipid metabolism in *Saccharomyces cerevisiae*. *Nature* **433,** 425–430 (2005).
53. Klipp, E., Nordlander, B., Kruger, R., Gennemark, P. & Hohmann, S. Integrative model of the response of yeast to osmotic shock. *Nature Biotechnol.* **23,** 975–982 (2005).
54. Abouhamad, W. N. *et al.* Computer-aided resolution of an experimental paradox in bacterial chemotaxis. *J. Bacteriol.* **180,** 3757–3764 (1998).
55. Kalir, S. & Alon, U. Using a quantitative blueprint to reprogram the dynamics of the flagella gene network. *Cell* **117,** 713–720 (2004).
56. Wiley, H. S., Shvartsman, S. Y. & Lauffenburger, D. A. Computational modeling of the EGF-receptor system: a paradigm for systems biology. *Trends Cell Biol.* **13,** 43–50 (2003).
57. Sasagawa, S., Ozaki, Y., Fujita, K. & Kuroda, S. Prediction and validation of the distinct dynamics of transient and sustained ERK activation. *Nature Cell Biol.* **7,** 365–373 (2005).
58. Martin, B. R., Giepmans, B. N., Adams, S. R. & Tsien, R. Y. Mammalian cell-based optimization of the biarsenical-binding tetracysteine motif for improved fluorescence and affinity. *Nature Biotechnol.* **23,** 1308–1314 (2005).
59. Hasty, J., McMillen, D. & Collins, J. J. Engineered gene circuits. *Nature* **420,** 224–230 (2002).
60. Levine, M. & Davidson, E. H. Gene regulatory networks for development. *Proc. Natl Acad. Sci. USA* **102,** 4936–4942 (2005).
61. Ortega, F., Acerenza, L., Westerhoff, H. V., Mas, F. & Cascante, M. Product dependence and bifunctionality compromise the ultrasensitivity of signal transduction cascades. *Proc. Natl Acad. Sci. USA* **99,** 1170–1175 (2002).
62. Milo, R. *et al.* Network motifs: simple building blocks of complex networks. *Science* **298,** 824–827 (2002).
63. Mangan, S., Zaslaver, A. & Alon, U. The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *J. Mol. Biol.* **334,** 197–204 (2003).
64. Schoning, J. C. & Staiger, D. At the pulse of time: protein interactions determine the pace of circadian clocks. *FEBS Lett.* **579,** 3246–3252 (2005).
65. Atkinson, M. R., Savageau, M. A., Myers, J. T. & Ninfa, A. J. Development of genetic circuitry exhibiting toggle switch or oscillatory behaviour in *Escherichia coli*. *Cell* **113,** 597–607 (2003).
66. Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in *Escherichia coli*. *Nature* **403,** 339–342 (2000).
67. Kramer, B. P. *et al.* An engineered epigenetic transgene switch in mammalian cells. *Nature Biotechnol.* **22,** 867–870 (2004).
68. Becskei, A., Seraphin, B. & Serrano, L. Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. *EMBO J.* **20,** 2528–2535 (2001).
69. Isaacs, F. J., Hasty, J., Cantor, C. R. & Collins, J. J. Prediction and measurement of an autoregulatory genetic module. *Proc. Natl Acad. Sci. USA* **100,** 7714–7719 (2003).
70. Kobayashi, H. *et al.* Programmable cells: interfacing natural and engineered gene networks. *Proc. Natl Acad. Sci. USA* **101,** 8414–8419 (2004).
71. Elowitz, M. B. & Leibler, S. A synthetic oscillatory network of transcriptional regulators. *Nature* **403,** 335–338 (2000).
72. Fung, E. *et al.* A synthetic gene-metabolic oscillator. *Nature* **435,** 118–122 (2005).
73. Hasty, J., Dolnik, M., Rottschafer, V. & Collins, J. J. Synthetic gene network for entraining and amplifying cellular oscillations. *Phys. Rev. Lett.* **88,** 148101 (2002).
74. Bulter, T. *et al.* Design of artificial cell–cell communication using gene and metabolic networks. *Proc. Natl Acad. Sci. USA* **101,** 2299–2304 (2004).
75. You, L., Cox, R. S. III, Weiss, R. & Arnold, F. H. Programmed population control by cell–cell communication and regulated killing. *Nature* **428,** 868–871 (2004).
76. Basu, S., Gerchman, Y., Collins, C. H., Arnold, F. H. & Weiss, R. A synthetic multicellular system for programmed pattern formation. *Nature* **434,** 1130–1134 (2005).
77. Jaenecke, S., de Lorenzo, V., Timmis, K. N. & Diaz, E. A stringently controlled expression system for analysing lateral gene transfer between bacteria. *Mol. Microbiol.* **21,** 293–300 (1996).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Reich, J. J. Collins, I. Mattaj and J. Gagneur for critical reading of the manuscript, and P. Beltrao and M. Isalan for helpful discussions. This work was financed partly by the EC grant, COMBIO.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to L.S. (serrano@embl.de).