

Modelling cellular behaviour

Drew Endy & Roger Brent

Representations of cellular processes that can be used to compute their future behaviour would be of general scientific and practical value. But past attempts to construct such representations have been disappointing. This is now changing. Increases in biological understanding combined with advances in computational methods and in computer power make it possible to foresee construction of useful and predictive simulations of cellular processes.

In molecular, cellular and developmental biology, compact and elegant theories of the sort familiar in physics are rare; rather, explanations of phenomena are typically couched in natural language narratives that describe the interactions of large numbers of distinct molecular entities. In this essay, we define *model* as any representation of a system. Models are usually made up of abstractions that are easier to manipulate than the actual system. We are concerned with models of cellular processes whose internal descriptions match the molecular mechanisms by which those processes act. In particular, we are interested in models that incorporate knowable quantities, including the number or amount of different entities (for example, proteins, transcripts or regulatory sites) and the rates at which these entities react, and in which the entities and reactions are governed by physical laws. We define *simulation* as a representation that embodies information contained in a model, and that provides access to the model by allowing computation of system behaviour.

To give an example of this usage, physicists routinely use computer simulations to access and elaborate predictions of the dominant theoretical framework in high-energy physics, the Standard Model. In the biological examples we will discuss here, the information that constitutes a model might be described in words or systems of equations, but the simulations that provide access to the models will run on computers.

The models used in molecular, cellular and developmental biology are typically heuristic. They arise alongside the process of experiment and are inseparable from it. Because they are based on experiments in which perturbation of system components has had observed effects, the models typically contain embedded knowledge of causality and of the passage of time. In such models, time progresses from one experimentally defined causal step to the next (Fig. 1a, b). In the simulations discussed here, time is absolute (Fig. 1c).

Too ambitious, too soon

Past efforts to model behaviour of molecular and cellular systems over absolute time

typically were qualitatively incomplete or oversimplified compared to available knowledge, and quantitatively incomplete in the sense that key numbers were unknown. For example, even thoughtful, carefully constructed models posited the control of embryonic and somatic cell proliferation by a single cyclin whose degradation controlled entry into mitosis¹ after the existence of different cyclins that controlled progression through different phases of the cell cycle was established². Models of circadian rhythms based on known molecular entities³ were immediately outgrown as new molecules (for example, Clock and Cycle) were discovered⁴. In general, such models did not result in predictions of phenomena that biologists perceived to be significant enough to warrant subsequent experimental effort.

An extreme example of the disjunction between model and experiment is the study of imagined networks of mutually activating and repressing genes (or gene products) — so-called ‘genetic regulatory networks’. Early studies showed that relatively simple interactions among network members could give rise to surprisingly complicated behaviour (ref. 5 and Fig. 2). However, by the early 1970s it was becoming apparent that few if any living systems had complex regulatory networks of this type, and that living systems regulate their transitions from state to state in other ways (see below). Although research on these imagined networks continues to this day, most biologists are either unaware of the work or ignore it.

But despite this history, we can now contemplate models of molecular, cellular and developmental biological systems that are coupled to experiment and result in increased understanding. One reason for optimism is that for some processes, enough biology is now known to begin to constrain useful models, and we can foresee obtaining much of the rest.

Qualitative simulations

One computable representation we may shortly expect to see is the so-called Biological Information System (BIS). The term

comes by analogy to Geographical Information Systems (GISs). BISs will extend current databases by embodying largely qualitative mechanistic knowledge. Over the next decade, BISs are likely to develop further, to encompass all known qualitative facts, including the components, their interactions and causal relationships for entire cellular subsystems and cells. The qualitative relationships among the components of such systems would be described by natural language equivalents — a small group of verbs defining permitted interactions. The information contained in BISs will be used to compute qualitative system behaviour over small numbers of causal steps. Although such computations will be only simple manipulations of existing knowledge, they will still be useful (see below).

Quantitative models

Progress in computation

Quantitative models of cellular processes often involve the representation of chemical reactions for which reactant molecules are scarce, and the continuous-variation approximation of differential calculus breaks down. Whereas in the 1950s the advent of the digital computer allowed numerical solution of large systems of differential equations, it was not until the 1970s that stochastic methods⁶ were developed to handle scarce reactants. During the 1990s these methods began to be applied widely to simulate biological systems^{7,8}. Recently, the efficiency of these methods has been increased significantly⁹, so that the earlier simulations⁸ can now be solved on desktop machines instead of supercomputers. Further reductions in computational cost will come from ‘linking’ deterministic and stochastic regimes (ref. 10 and D. T. Gillespie, personal communication), and may come from new methods that better handle large numbers of coupled reactions.

The promise of these methods also depends on increases in computing power. For example, one can now use a Gibson-modified Gillespie algorithm⁹ to execute 10¹⁰ reaction events per day on an 800-MHz

than warm ones¹⁹. Other applications of physics are less obvious. For instance, particular models of circadian rhythms are sensitive to noise and, given variation in the timing of reaction events, do not produce the stable oscillations observed in nature²⁰. Thus, for circadian rhythms, consideration of system function and likely process physics helped to dismiss a particular class of models. Looking beyond these cases, it is interesting to note that much of cell and organismic biology can be understood as the processing of information from the genome, from internal events, and from external events, by an amorphous 'architecture' of diffusing molecular components. We can thus hope that future developments in information theory will provide broader insights into biological function and help constrain models and suggest experiments.

Finally and most importantly, we need to devise new experimental methods for obtaining quantitative data about biological processes. At the moment, we lack good experimental means to determine: (1) the absolute numbers of different molecular species in populations of cells; (2) the numbers of these species in individual cells; (3) how those numbers vary among individual, genetically identical members of a population; (4) how those numbers vary over time; and (5) the rates of the individual reactions causing that variation. The development of methods for acquiring this quantitative knowledge is one of the greatest challenges for biology in the twenty-first century²¹, one well beyond the scope of this essay.

Need to choose useful levels of resolution

Any model embodies a physical and logical level of resolution. It seems likely that for many cellular and early embryonic developmental processes, the appropriate level of resolution is that of known proteins, DNA regulatory sites, and so on. However, in any given instance, an assumption that those molecules are 'localized' to well-mixed compartments may not be sufficient. For example, the discovery that *E. coli* MinC and MinD, proteins that suppress septum formation and cytokinesis, are localized to the poles, helped explain why the cell normally divides in the middle^{22,23}. But the same experiments revealed the startling fact that individual protein molecules do not stay put. Molecules of MinC and MinD translocate from one pole to the other over tens of seconds. The use of fluorescent fusion proteins (and other methods) will surely reveal many instances where spatial localization is important for understanding process function. It is thus likely that future simulations will need to divide the cellular milieu into individual voxels (volume elements) in which reactions occur.

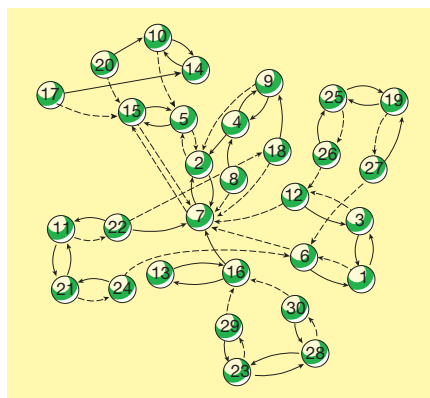


Figure 2 Behaviour of a 'genetic regulatory network'. In this example, regulation comes from cross-acting activators and repressors, and each of the several hundred genes is regulated by the products of two others. Shown are transitions among the 30 stable states of this network⁵. State transitions in such networks show, for example, basins of attraction and chaotic regimes⁴². In general, living systems seem to use other mechanisms to regulate their transitions from state to state (see text).

Future simulations will also need to allow for transition among different levels of resolution. A biologist might describe a protein as a simple ellipsoid, then, in the next breath, explain the effect of a point mutation by the atomic-level structural changes it causes in the active site. We can imagine a future simulation of an intracellular signalling pathway that ignored the shape and size of the individual molecular components, except when computing the effect of a kinase inhibitor when specific atomic information would be required about the interaction of the inhibitor with an active site. Similarly, future simulations of a cell (or groups of cells) might treat individual signal transduction pathways as parameterized modules, until pathway-specific effects needed to be represented. By allowing transitions to the coarsest level of resolution needed to represent observed behaviour, future simulations will use fewer computer cycles, and facilitate the ability of researchers to comprehend and interact with them.

Need to interact with experiment

Just as biological models were developed through the comparison of model-based predictions with experimental observations, so simulations of biological systems will need to develop alongside of, and in comparison with, experiments. The level of comparison will sometimes be qualitative. For example, a simulation of T7 growth²⁴ allowed the prediction that some rearranged genomes should encode phage that grow faster than wild type (a prediction that for the single genome tested to

date proved incorrect²⁵). At other times the level of comparison will be quantitative, based on static endpoints. For example, Kananyan *et al.*²⁶ and Arkin *et al.*⁸ compared the computed number of λ phage that form lysogens as a function of multiplicity of infection to experimental observations of Kourilsky *et al.*²⁷. In the future, the important level of comparison may frequently be quantitative, based on time-dependent behaviour. For example, discrepancies between computed and observed phage T7 protein synthesis rates suggested that translation from some T7 messenger RNAs might be subject to negative regulation by an as-yet-unknown mechanism (Fig. 3 and ref. 25).

There are very few biological systems for which complete quantitative models can be constructed from existing information. Because contemporary biologists have no shortage of hypotheses they find worth pursuing, virtually all generated without recourse to quantitative models, the information needed to construct them will not automatically be forthcoming. Thus, to be successful, future modelling efforts will probably need to direct and influence ongoing experiments.

Need to define model systems

Sometimes it will be easier to gather experimental data from simplified systems. Consider the previously mentioned yeast signal cascade. If there are 25,666 unique protein complexes that contain Ste5, and it is unknown which occur *in vivo*, and experimental determination of complex existence is not easy, we might reduce the number of unique complexes by fusing individual protein monomers into chimaeras that retain biological function (P. M. Pryciak, personal communication). Simplified systems can even be constructed from scratch. For example, several groups have constructed simple genetic systems using prokaryotic repressors. So far, the synthetic systems constructed have been relatively simple, with around a dozen genetic components^{28,29}. However, as the ability to synthesize and assemble large DNA fragments³⁰ continues to increase (and the cost of synthesis decreases), more ambitious systems will be designed and constructed.

Still, a good deal of information for quantitative models will be gathered from non-simplified systems — cells and organisms. The organisms and cell types may or may not be well studied. But in this genomic age, if it seems appropriate to develop a hitherto understudied organism into an experimental system, we can at least hope to bring about a basic level of understanding by marshalling the full power of sequencing, gene expression monitoring, large-scale mapping of protein interactions, functional

insight feature

analysis by transposon mutagenesis, deletion mutagenesis, and dominant protein-based approaches²¹.

Control of behaviour by genes

Figure 4 shows an eighteenth-century orrery, a quantitative simulation of the motion of the planets in the Solar System. Although the observations on which the simulation was based and the understanding of the physical laws that governed its elements were, in retrospect, quite accurate, the computed positions of the planets eventually deviate from what is observed. Deviation from observation is due both to imperfections in the clockwork, the brass rings and gears, and to the fact that, over long periods of time, the motion of the planets around the Sun is chaotic³¹. Although perhaps not chaotic, in biological systems (and simulations), too much depends on chance interactions among small numbers of interacting molecules to yield behaviour that is completely determined over time.

However, aspiring modellers can make use of the fact that cells and organisms use a number of genetic mechanisms to supplement their highly imperfect biochemical clockwork and keep their dynamic behaviour on track. First, biological systems frequently go back to the genome, invoking subprogrammes that reset them into new states. Regulatory proteins, frequently gene activators, initiate these genomic subprogrammes. For example, expression of MyoD protein initiates a course of gene expression that converts fibroblasts into myoblasts (muscle precursors)³². Similarly, ectopic expression of the Eyeless protein in the future leg, wing or antenna tissues of developing *Drosophila melanogaster* larvae invokes a subprogramme that results in (nonfunctional) eyes at the sites of Eyeless expression³³. Once initiated, progression through any given process may rely on biochemical clockwork. But at some point the subprogramme is completed, and progression to the next process presumably requires invoking a new subprogramme.

Second, cells use checkpoint controls — feedback mechanisms that prevent a sequence of events from starting, and hold the cell at the ‘checkpoint’ until the mechanism receives a signal that a required sequence of earlier events has in fact been completed. Checkpoints are defined operationally, for example by mutations that arrest progression of cellular systems at given states. For example, in the yeast *Saccharomyces cerevisiae*, the checkpoint protein Rad9 prevents cells with DNA damage from attempting a new round of DNA synthesis until the damage is repaired³⁴. Both genomic subprogrammes and checkpoint controls punctuate the temporal transitions of systems and provide the opportunity to reset them to new starting states.

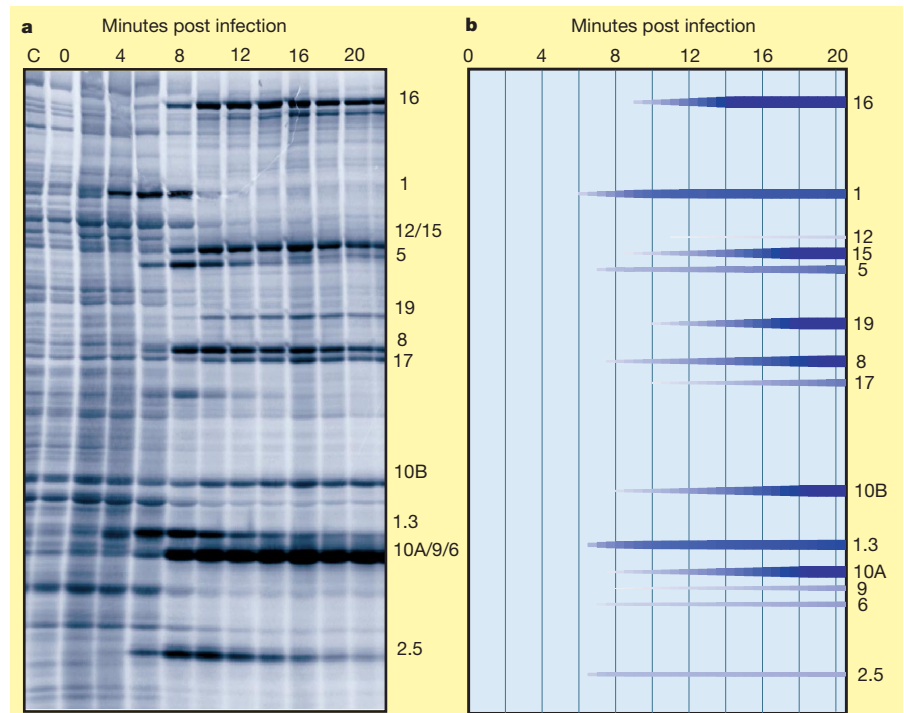


Figure 3 Observed and computed rates of phage T7 protein synthesis. **a**, Experimental determination of time and amount of phage protein synthesis during T7 infection (see <http://virus.molsci.org> for experiment and simulation details). **b**, Simulation output. The T7 simulation is based on current biological knowledge. Note that comparison of observed and computed T7 protein synthesis rates reveals that during an actual phage infection, synthesis of the T7 proteins gp1, gp1.3, gp2.5 and gp5 is negatively regulated by unknown mechanism(s).

Third, cells and organisms use other, less well understood mechanisms that coordinate timing of biological events and place dynamic system behaviour under more regulation than could be provided by biochemical clockwork alone. For example, *clk-1* mutants of the nematode *Caenorhabditis elegans* develop slowly at 15 °C, faster at 20 °C, and still faster at 25 °C. When two-cell *clk-1* embryos removed from 15 °C mothers are shifted to 20 °C, they continue to develop slowly, whereas two-cell embryos from 25 °C mothers shifted to 20 °C continue to develop rapidly^{35,36}. This observation shows that — beginning at a developmental stage before transcription of the embryo’s own genes starts — the tempo of development in the wild-type worm is specified by a mechanism that is in part temperature independent. Construction of quantitative models can only further focus experimental attention on *clk-1* and other mechanisms that govern the timing of biological processes. Dedicated mutant hunts and ‘protein genetic’ screens³⁷ may reveal additional ways by which cells and organisms coordinate and regulate their time-dependent behaviour and reset to new states.

Consequences of success

The increasing amount of biological knowledge will probably itself be sufficient to force

the development of BISs to contain it. Such information systems will be good for more than teaching and learning. Relatively simple operations on (partly quantitative) information in GISs now allow people to determine driving directions and distance. By analogy, consider an information system that embodies known interactions and causal relationships among proteins that regulate cell division, and which could use that knowledge to enumerate those entities affected by perturbing the activity of different members of the protein network. Imagine a physician performing cancer therapy in 2020 who is looking at a listing of the changes in DNA sequence, gene expression and proteins in an individual tumour. The physician might use this information together with a BIS to support decisions on whether the inhibition of a particular protein kinase is likely to be useful for treating that particular tumour.

Vetting information

Another consequence would be an increase in the accuracy of biological information. This increase arises naturally from the fact that large-scale modelling efforts will require the combining of information from many different sources. Biology currently tests the validity of qualitative conclusions from different laboratories by mechanisms that range from peer-review to gossip. These are

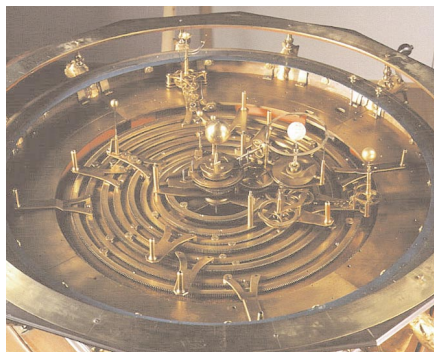


Figure 4 A simulation from 1773. Figure shows the workings of the 'Grand Orrery', a mechanical device that computes the positions of planets and moons in the Solar System (see <http://www.nmsi.ac.uk/collections/exhiblets/george3/gallery.htm>). Less than 80 years later, models of planetary orbits were precise enough to demonstrate that deviations in the path of Uranus from its expected orbit could be accounted for by positing the existence of a new planet, and to tell astronomers where to point their telescopes to find it. Thus, by the 1840s, astronomical simulations were precise enough to allow prediction of Neptune. By contrast, during the 1990s no biological model of circadian rhythm allowed prediction of the regulatory proteins *Cycle* or *Clock*.

fairly effective; for example, they were able to demonstrate that not all 'phage T7 labs' were actually studying T7 (ref. 38). However, agreement on sets of quantitative information (and on very large sets of qualitative information) will probably require new ways of checking the accuracy, consistency and validity of that information. Making the computable information, the models and the simulations available to all scientists is clearly part of the solution. Once a draft simulation is constructed, discrepancies between computed and observed system behaviour will suggest changes to the model and new experiments. Done properly, providing access to simulations to large communities of biologists should accelerate the process of biological discovery itself.

Guiding intervention and therapy

Another consequence of success comes from the fact that quantitative mechanism-based models allow researchers to observe the complete behaviour of a specified system over time, and track all changes in its behaviour due to perturbations. It is easier to use a model to search for perturbations that have significant effects on system behaviour than it is to perform similar experiments on the living system. In some systems, an experimental search for sensitive components may not be possible. Moreover, models allow the search for multiple small perturbations that produce large effects when combined. In

most experimental systems, this is usually not possible.

Such capabilities will be useful for drug discovery and therapy. For example, quantitative models would help identify target proteins that give rise to therapeutic effects when partially inhibited. This alone would allow the development of small-molecule inhibitors that bind the target protein less tightly, thereby reducing the time needed to discover new drugs. Even more benefit may come from identification of cases where large changes in system behaviour could be achieved by partial inhibition of multiple protein targets. This would allow the identification of multiple targets that would permit the use of two or more drugs in smaller amounts, potentially resulting in fewer side effects. Models may also be useful in regimes (for example, anticancer therapy) in which drug concentration or amount of inhibition is limited. For example, models have been used to indicate that inhibition of a particular 'drug target', gene 1 messenger RNA of phage T7, has a paradoxical effect. The encoded protein, gp1, downregulates its own activity. Mutations in the messenger RNA that decrease 'drug' binding result in greater system inhibition³⁹.

Improving biological design

Models should form the basis of tools to aid in optimization of existing biological systems and design of new ones. Additionally, quantitative models will enable engineers to evolve biological systems by rounds of variation and selection for any function they desire. Such model-based evolution may complement existing organismic (for example, crossing two strains) and molecular (for example, mutagenesis using polymerase chain reaction, or DNA shuffling) approaches^{40,41} that depend on sometimes clever but often cumbersome selections and screens in the real world. As mentioned above, by the time computer-based optimization of living systems is possible, it will also be possible to fabricate large DNA sequences encoding the successful solutions, and thus to transfer successful designs from model to life.

Enabling new scientific understanding

Finally, mechanism-based models may bring now-unforeseen benefits to scientific understanding and capability. We have hinted at three of these. One comes from the fact that current biological understanding (and experimental methodology) does not deal very well with the passage of absolute time. The experiments needed to construct quantitative models, and consideration of those models, may help reveal mechanisms and insights into ways living systems regulate their temporal behaviour. A second comes from the idea that many biological systems can be described in terms of information processing. Quantitative models will be

needed to gain any insights from this metaphor. A third comes from the fact that mechanism-based models will be used as design tools and should speed the rise of a greatly heightened capability to engineer living systems. Although the lineaments of a world in which biology is directed by human intention might be foreseeable, the details of the changes to our selves and to our interaction with the living world cannot be foreseen.

Drew Endy and Roger Brent are at the Molecular Sciences Institute, 2168 Shattuck Avenue, Berkeley, California 94704, USA.

1. Tyson, J. J. *Proc. Natl Acad. Sci. USA* **88**, 7328–7332 (1991).
2. Lehner, C. F. & O'Farrell, P. H. *Cell* **61**, 535–547 (1990).
3. Leloup, J. C. & Goldbeter, A. *J. Biol. Rhythms* **13**, 70–87 (1998).
4. Rutila, J. E. *et al. Cell* **93**, 805–814. (1998).
5. Kauffman, S. A. *J. Theor. Biol.* **22**, 437–467 (1969).
6. Gillespie, D. T. *J. Comput. Phys.* **22**, 403–434 (1976).
7. McAdams, H. H. & Arkin, A. P. *Proc. Natl Acad. Sci. USA* **94**, 814–819 (1997).
8. Arkin, A., Ross, J. & McAdams, H. H. *Genetics* **149**, 1633–1648 (1998).
9. Gibson, M. A. & Bruck, J. J. *Phys. Chem.* **2104**, 1876–1889 (2000).
10. Gillespie, D. T. *J. Chem. Phys.* **113**, 297–306 (2000).
11. van der Steen, A. J. & Dongara, J. J. <www.top500.org/ORSC/> (2000).
12. Hutcheson, G. D. & Hutcheson, J. D. *Sci. Am.* 54–62 (January 1996).
13. Cayley, S., Lewis, B. A., Guttman, H. J. & Record, M. T. *J. Mol. Biol.* **222**, 281–300 (1991).
14. Zimmerman, S. B. & Trach, S. O. *J. Mol. Biol.* **222**, 599–620 (1991).
15. Elowitz, M. B., Surette, M. G., Wolf, P.-E., Stock, J. B. & Leibler, S. *J. Bacteriol.* **181**, 197–203 (1999).
16. Bardwell, L., Cook, J. G., Chang, E. C., Cairns, B. R. & Thorner, J. *Mol. Cell. Biol.* **16**, 3637–3650 (1996).
17. Morton-Firth, C. J. & Bray, D. *J. Theor. Biol.* **192**, 117–128 (1998).
18. Block, S. M., Segall, J. E. & Berg, H. C. *Cell* **31**, 215–26 (1982).
19. Aho, A.-C., Donner, K., Hyden, C., Larsen, L. O. & Reuter, T. *Nature* **334**, 348–350 (1988).
20. Barkai, N. & Leibler, S. *Nature* **403**, 267–268 (2000).
21. Brent, R. *Cell* **100**, 169–183 (2000).
22. Hu, Z. & Lutkenhaus, J. *Mol. Microbiol.* **34**, 82–90 (1999).
23. Raskin, D. M. & de Boer, P. A. J. *Bacteriol.* **181**, 6419–6424 (1999).
24. Endy, D., Kong, D. & Yin, J. *Biotechnol. Bioeng.* **55**, 375–389 (1997).
25. Endy, D., Yu, L., Yin, J. & Molineaux, I. J. *Proc. Natl Acad. Sci. USA* **97**, 5375–5380 (2000).
26. Kanayan, G. Kh., Ratner, V. A. & Churavaev, R. N. *Genetika* **16**, 2209–2017 (1980).
27. Kourilsky, P. *Mol. Gen. Genet.* **122**, 183–195 (1973).
28. Elowitz, M. B. & Leibler, S. *Nature* **403**, 335–338 (2000).
29. Gardner, T., Cantor, C. R. & Collins, J. J. *Nature* **403**, 339–342 (2000).
30. Yount, B., Curtis, K. M. & Baric, R. S. *J. Virol.* **74**, 16000–10611 (2000).
31. Sussman, G. J. & Wisdom, J. *Science* **257**, 56–62 (1992).
32. Davis, R. L., Weintraub, H. & Lassar, A. B. *Cell* **51**, 987–1000 (1987).
33. Halder, G., Callaerts, P. & Gehring, W. J. *Science* **267**, 1788–1792 (1995).
34. Weinart, T. A. & Hartwell, L. H. *Science* **241**, 317–322 (1988).
35. Wong, A., Boutis, P. & Hekimi, S. *Genetics* **139**, 1247–1259 (1995).
36. Branicky, R., Benard, C. & Hekimi, S. *BioEssays* **22**, 48–56 (2000).
37. Colman-Lerner, A. & Brent, R. *Trends Cell Biol.* 56–60 (Suppl. December 2000).
38. Studier, F. W. *Virology* **95**, 70–84 (1979).
39. Endy, D. & Yin, J. *Antimicrob. Agents Chemother.* **44**, 1097–1099 (2000).
40. Soong *et al. Nature Genet.* **25**, 436–439 (2000).
41. Schmidt-Dannert, C., Umemo, D. & Arnold, F. H. *Nature Biotechnol.* **18**, 750–753 (2000).
42. Kauffman, S. A. *The Origins of Order: Self-Organization and Selection in Evolution* (Oxford Univ. Press, 1993).

Acknowledgements. We thank R. Carlson, A. Colman-Lerner, D. Gillespie, M. Gruber, P. Pryciak, C. Kenyon, E. Kroll, E. Lyons, L. Lok, I. J. Molineux, O. Resnekov, T. Roosevelt, L. Thomason, J. Yin and L. You for useful comments, discussions or unpublished information. Work at TMSI is supported by grants to R.B. and D.E. from the NIH, DARPA and the Office of Naval Research.