EMBO *reports*

The evolution of biology

A shift towards the engineering of prediction-generating tools and away from traditional research practice

Lawrence Kelley & Michael Scott

One of the most interesting recent trends in the biosciences has been the development of research methods that do not adhere to familiar standards of scientific practice. The traditional aim of scientific research—most notably in physics—has been to gain a comprehensive understanding of natural phenomena and to generate hypotheses that provide simple, law-like and broad explanations. Contemporary research in bioinformatics is markedly different: bioinformaticians are increasingly generating tools to make accurate predictions for a restricted range of phenomena, irrespective of their simplicity or broader application in science. The new developments within the biological sciences are also hard to square with traditional *a priori* theories—inductivism, falsificationism or inference to the best explanation—that serve to articulate the standards of proper scientific methodology.

This shift towards the engineering of prediction-generating tools—and away from traditional research practice and the creation of cognitively accessible explanations—shows that we need to rethink what embodies respectable scientific practice. The ideas of simplicity, law-likeness and an ability to understand what it is that makes nature 'work', are widely conceived—both within and outside science—not merely as a desirable product of scientific investigation, but more so as the characteristics of science that distinguish it from fields such as theology or literature. In fact, the implied or explicit promise of fully understanding natural phenomena is not only a standard component of most research projects, but also a major selling point of research that attracts funding.

Consequently, there is a widespread and continuing expectation that the biological sciences could and should provide theoretical insight into the workings of biological phenomena, despite a growing body of evidence to the contrary. This attitude is arguably sustained, at least in part, by the successful elucidation of the structure of DNA by Francis Crick (1916–2004) and James Watson. Their discovery, which came at an early stage in the development of modern biology, has given undue credence to an optimistic but unfounded belief that further research in the biological sciences will uncover similarly fundamental and simple insights.

But biological results and methodologies, particularly as they have increased in predictive power, look entirely unlike the theoretical approaches of physics. The main reason for this development—as we point out here—is the enormous complexity of biological organisms, which defy any attempt to explain or understand them using simple hypotheses or mathematical rules. There is a growing awareness that this complexity presents a considerable challenge to the reductionist approach in biology (Mazzocchi, 2008; Gannon, 2007). Here, we focus on the even more dramatic implications for scientific methodology and offer a radically alternative way of thinking about 'good' science, which is informed by evolutionary considerations.

The inductivist theory of the scientific method holds that science should proceed by selecting those theories that are best confirmed by past observations. A standard problem with this method, which was proposed by the Austrian philosopher Karl Popper (1902–1994), is that most scientific theories make generalizations, based on a finite number of observations, which have an infinite number of possible instantiations (Popper, 1963). No matter how often we put a theory to the test, we will never be able to verify its truth for a significant proportion of all possible instantiations. Consequently, scientific theories have the status of conjectures: we might be able to prove them false, but we have no reason to think that they are true, or even probably true.

> **…there is a widespread and continuing expectation that the biological sciences could and should provide theoretical insight into the workings of biological phenomena, despite a growing body of evidence to the contrary**

The Israeli physicist David Deutsch derided the inductivist approach because it fails to appreciate the essentially problem-solving nature of scientific inquiry (Deutsch, 1998). Instead, he argued, science progresses by conducting experiments, generating conjectures, and replacing them if they turn out to be false. However, despite its successful characterization of some aspects of scientific practice—at least in physics—Deutsch's falsificationism should have little appeal for the biological sciences. One problem is that it runs into trouble with probabilistic hypotheses. Any hypothesis that states "all Xs are Ys" can be falsified with just one observation that an X is not a Y. But, if the hypothesis states that "90% of Xs are Ys", how many observations do we need—and how large should the sample size be—before we can safely conclude that the conjecture is false? If the best hypothesis corresponds to a statistical relationship discovered between Xs and Ys in what scientists deem to be a sufficiently large sample, then—contrary to Deutsch's preferred method—we are reasoning from induction.

The main problem in applying either inductivism or falsificationism—and other related *a priori* proposals—to the biological sciences is actually neither technical nor philosophical. Rather, by placing general theories at the centre of the scientific method, these proposals largely fail to accord with current scientific practice. If the standards required by these methods were taken seriously, then much research in the biological sciences would be deemed unscientific, and what remained would be far less effective at solving problems. We demonstrate this point by describing the development of research strategies in various fields of biological research.

Consider, for example, the fields of genomics and proteomics: the past years have seen an enormous increase in the amount of experimental data available, which is due to whole-genome sequencing, high-throughput structural genomics, microarray technology and protein–protein interaction assays. More than seven million protein sequences are now known and more than 50,000 protein structures have been experimentally determined. This enormous wealth of data requires the extensive use of computer tools for storage, retrieval and, most relevantly, analysis and prediction.

> **The planets and comets in our solar system all abide by easily understood and simple principles, so why should the same not be true for protein folding?**

The three-dimensional structure of a protein is a crucial determinant of its biological function. Unfortunately, the current experimental techniques to determine the structure of a protein are expensive, time-consuming and sometimes do not produce any results at all. This is particularly true in the case of membrane-spanning proteins and explains why the number of experimentally determined protein structures is two orders of magnitude lower than the number of unique protein sequences. To overcome this bottleneck, biologists and computer experts have been developing methods to predict the structure of a protein from its sequence.

This is not an easy undertaking. A typical protein could theoretically adopt between 1×10^{100} and 1×10^{500} conformations. The folding of a protein into its final and active structure is a complex, nonlinear and dynamic interaction between the individual amino acids along its length and the surrounding solvent—a process that involves thousands of atoms interacting with thousands of water molecules among others *in vivo*. Every part of a sequence can interact with every other part through a complex combination of electrostatic forces, steric effects and hydrogen bonding. The basic principles that drive protein folding are reasonably well understood at the level of physics and chemistry, but the size and complexity of proteins makes conventional simulation impossible; indeed, more than 20 years of intense research efforts and ever more sophisticated computer simulations have found no straightforward connection between the amino-acid sequence of a protein and its tertiary structure.

So far, the only reasonably successful methods for predicting the structure of a protein make use of empirical and statistical models, machine learning, evolutionary data-mining and the statistical sampling of conformational space. These methods are becoming constantly more sophisticated and more successful at accurately predicting a structure, and theoreticians and experimentalists alike routinely use their predictions. However, the main objective of these methods is neither to gain an understanding of the folding mechanisms, nor to extract a simple or law-like hypothesis about the relationship between the structure of a protein and its amino-acid sequence, but instead to generate the most effective predictor of that relationship. The predictions themselves are, of course, open to experimental verification. But testing the predictions does not verify or falsify any theory about how protein folding works; it only measures the success of the particular project or accuracy of the software that has yielded the predictions.

Making predictions based on the available data without any broader theoretical understanding is different to generating hypotheses about underlying mechanisms, and it is therefore a significant step away from the traditional scientific methodology. One could respond that this apparent shift is due to technological problems such as the cost and length of time required to experimentally determine protein structures. After all, why should biological systems be fundamentally different to solar systems, for example, which consist of a huge number of different objects of varying size and structure? The planets and comets in our solar system all abide by easily understood and simple principles, so why should the same not be true for protein folding?

> **The best possible biological account of an evolved mechanism will probably be immensely messy and offer no cognitively appealing insight into its workings**

Another strategy to respond to this changing nature of biological research is to adopt more modest claims of what biological theories can explain. Some contemporary philosophers of science, who acknowledge the complexity of biological systems and the lack of universal, law-like principles that guide their behaviour, have argued that biologists explain phenomena by positing mechanisms. The behaviour of these mechanisms is predictable and subject to law-like principles, but with a narrower range of applications (Schaffner, 1993; Sarkar, 1998; Wimsatt, 1976; Machamer *et al*, 2000).

Of course, developing a theory or trying to discover a mechanism can still have a useful purpose in the biological sciences. For example, once technology has suitably advanced and more information on protein structures is available from which to draw general principles, it might be possible to start testing hypotheses about protein folding—although it is difficult to see what those hypotheses would look like. But, the general point is that although it is a good approach to posit conjectures for some problems—as it clearly was for great physicists such as Galileo Galilei (1564–1642), Isaac Newton (1643–1727) and Albert Einstein (1879–1955)—it is not usually the best strategy for solving biological problems. Moreover, there is a compelling argument to suggest that this strategy is not something that scientists should even aspire to when it comes to tackling certain problems.

Living organisms have evolved over long periods of time in response to changing environments that might have been entirely local to them or, if not local, might no longer exist. This statement underpins two salient points. First, there is a lack of comparative information about whether and how ancestral organisms that are radically different to the current inhabitants of the Earth could survive and proliferate in Earth-like environments. Moreover, we are not able to test whether current organisms would have evolved in the same way if one could 're-run' evolution under the same conditions (Service, 2007). Second, the survival of an evolving mechanism requires some level of fidelity in its reproduction, and robustness against environmental changes and errors in replication.

Yet, there is no *a priori* reason why these features should either require or be enhanced by conceptually simple mechanisms. This is particularly apparent when we study evolution at the molecular level, where random or pseudo-random variation coupled with non-random selection has driven the development of complex organisms. Random variations survive selection because they are phenotypically neutral, advantageous or insufficiently deleterious. Without any evidence that selection favours simplicity, an organism that has evolved through the accumulation of random changes will probably be complex. We therefore lack much of the information that would help us to work out why an organism functions as it does and why it evolved the way it did.

This might seem to be an unduly pessimistic outlook for the biological sciences given that we find complex physical structures that nonetheless exhibit law-like behaviour. For example, the patterns formed by grains of sand on a beach are highly complex, but the principles of erosion are relatively simple. In biology, too, there is one simple high-level generalization: the theory of evolution. However, this comparison between the analysis of complex systems in physics and biology quickly breaks down when we consider that, in physics, most structures exhibit simplicity at the level at which we want to understand their operation. No similar principles are available to explain the behaviour of many biological systems and structures.

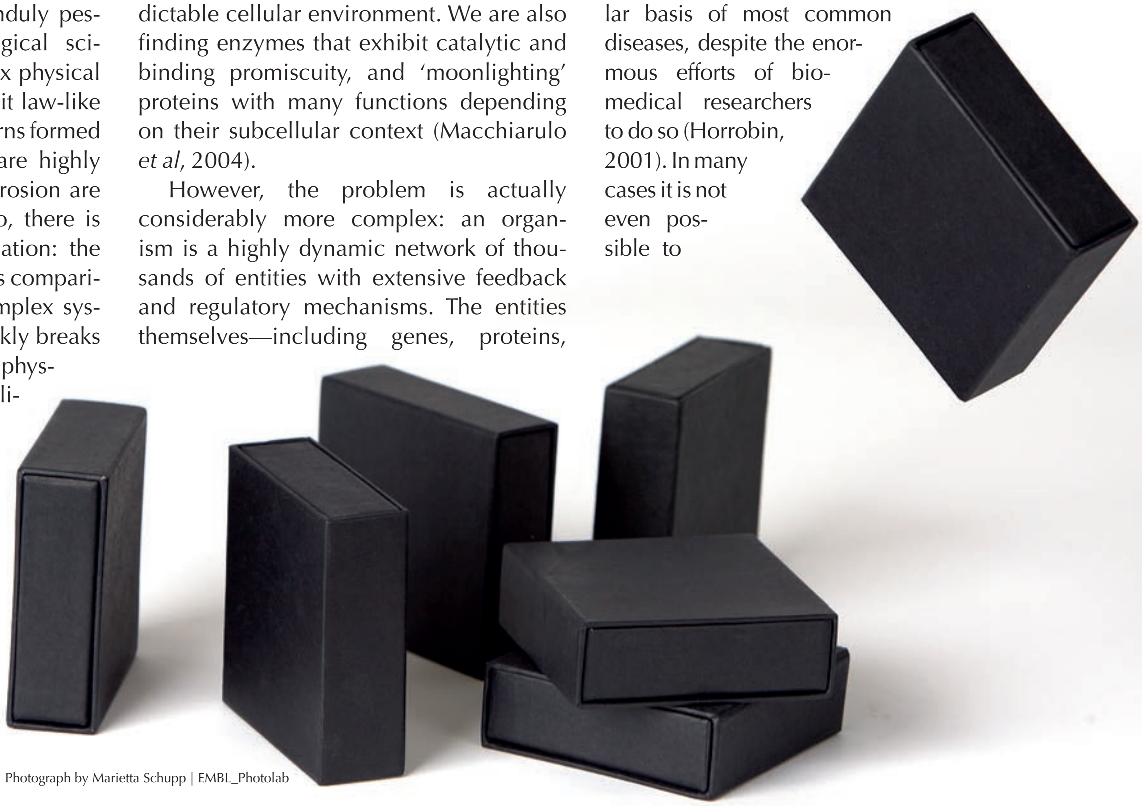
Taken together, these observations indicate that biologists should not expect evolved biological phenomena to be amenable to theoretical generalization above the low-level principles of chemistry and physics. The best possible biological account of an evolved mechanism will probably be immensely messy and offer no cognitively appealing insight into its workings. This conclusion is corroborated by a range of recent findings showing that, far from uncovering simple mechanisms at the foundations of biology, we find an astonishing degree of complexity.

Consider, for example, the simple hypothesis of "one gene, one protein, one function". This simplistic view was quickly overturned by the discoveries of alternative splicing, RNA interference, transposable elements and overlapping reading frames. We are also discovering that proteins can exist in an ensemble of various conformational states (Goh *et al*, 2004) and that splicing occurs frequently within protein domains (Birzele *et al*, 2008). Almost identical protein sequences can adopt completely different folds and functions (Alexander *et al*, 2007), whereas context-dependent 'chameleon' sequences can undergo radical changes in topology and function depending on their interactions with other proteins (Andreeva & Murzin, 2006). There is also a widespread recruitment of chaperone systems to guide folding (Hartl & Hayer-Hartl, 2002), in part to compensate for the complex and unpredictable cellular environment. We are also finding enzymes that exhibit catalytic and binding promiscuity, and 'moonlighting' proteins with many functions depending on their subcellular context (Macchiarulo *et al*, 2004).

However, the problem is actually considerably more complex: an organism is a highly dynamic network of thousands of entities with extensive feedback and regulatory mechanisms. The entities themselves—including genes, proteins, organelles and membranes—are dynamic ensembles of structure–function relationships. The complexity of biological signalling networks, for example, is determined by factors such as the number of components and the intricacy of the interfaces between them, the number and intricacy of conditional branches, or the degree of nesting and the types of data structure. In addition, there are the issues of dynamic assembly, translocation, degradation and the channelling of chemical reactions. All of these activities occur simultaneously and each component participates in several different activities (Weng *et al*, 1999).

Approaching the subject from an engineering viewpoint, Csete & Doyle (2002) argue that this enormous complexity is the result of an evolutionary trade-off between robustness, feedback and fragility. Instead, we claim that much of this complexity results from the randomness of the evolutionary process by which beneficial accidents are retained, regardless of their potential future shortcomings; evolution cannot see into the future. If computer software were designed in this way, the resulting programs would be complex and messy, albeit effective, as has been found in the field of evolutionary computing.

It is therefore not surprising that we are still unable to understand fully the molecular basis of most common diseases, despite the enormous efforts of biomedical researchers to do so (Horrobin, 2001). In many cases it is not even possible to



Photograph by Marietta Schupp | EMBL_Photolab see the phenotypic effect of a gene knockout (Pearson, 2002) simply because the genes acting in parallel pathways can compensate for missing genes or because the phenotype appears only under certain environmental conditions. We should also not be surprised at the "startling failure of the pharmaceutical research effort" (Horrobin, 2001), or of the growing doubts about the value of the reductionist approaches that focus on the specific genes or other molecular components of a biological system and overlook its more complex and systematic interactions (van Regenmortel, 2004; Kellenberger, 2004).

In summary, we see overwhelming complexity, and a mixture of structure and noise in most fundamental processes in molecular biology—transcription, translation and folding—through to the level of dynamic signalling and metabolic networks, and beyond that to the phenotype. These mechanisms exist simply because they work, or because they have worked in the past; sometimes, as with transposons, they might no longer have a crucial function. The result is that biologists have far too much unstructured information to be able to make informative conjectures—at a level higher than the standard chemistry or physics that guide specific interactions—with any prospect of being verifiable or true. Moreover, we also have too little information about the context in which these mechanisms evolved to know what a good conjecture might look like, and we have no grounds to believe that the workings of evolved mechanisms should be controlled by conceptually appealing principles.

> **It is therefore not the method that should control our conception of what counts as good science, but rather the measure of success**

Of course, our view of this problem could change with a dramatic new discovery—alien life, for example—but our current information and knowledge about evolutionary processes give us no reason to think that evolved mechanisms should be explicable at a higher level than the laws of physics and chemistry. And yet, notwithstanding the lack of simple explanations and theories, predictive models have proved to be remarkably effective.

Consider, for example, the use of predictive statistical and machine learning techniques. Gene-finding, modelling genetic networks, prediction of the secondary and tertiary structure of proteins, recognition of splice sites, protein–protein interactions, pattern recognition in microarray experiments, cancer diagnosis, tumour classification, motif discovery, drug design and phylogenetics are all examples in which researchers have developed computational techniques that make predictions without offering explanations of how the mechanisms under investigation function (Larranaga *et al*, 2006). All of these have made use of many techniques to analyse an otherwise insurmountable quantity of data, including neural networks, support vector machines, Bayes nets, statistical potentials, genetic algorithms, random forests, simulated annealing, Monte Carlo sampling and hidden Markov models.

Let us consider one example of how these predictive methods work: the use of supervised learning to find DNA sequences that are involved in the control of gene expression. Faced with a vast amount of experimental data in the absence of a general theory, it is useful to search for unusual patterns that might indicate some underlying structure. For example, we have experimental information that certain sequences are involved in the control of gene transcription, but we are often unable to see any clear patterns within such sequences that would differentiate them from other, non-functional sequences. We have, however, good reason to assume that there must be some pattern as the cellular machinery is able to correctly recognize control sequences.

Bioinformaticians commonly use a form of supervised machine learning in such cases. This involves collecting a training set of positive and negative examples—known transcription factor binding sites and non-binding sites, for example—and analysing them by using a wide range of algorithms to generate approximate functions that correctly classify the examples with greater or lesser accuracy. These 'learnt' functions can then be applied to new experimental data for automatic classification and prediction. Alternatively, unsupervised learning methods—such as Gibbs sampling—start out with data sets that are believed, on the basis of independent evidence, to contain a characteristic but unknown nucleotide pattern, which might represent a binding site (Tompa *et al*, 2005).

For many years, the machine learning community has been developing progressively more powerful algorithms. Unlike early attempts, most modern learning programs are not limited to simple linear functions but can also learn nonlinear functions in large-dimensional spaces with a high degree of accuracy and computational efficiency. However, the vast majority of such algorithms produce a 'black box' that often consists of hundreds or thousands of parameters in an abstract space, rather than a linear, logical instruction set; it is generally impossible to reverse-engineer these solutions into any meaningful rule or principle.

In the field of protein structure prediction, the standard practice is to use a dynamic data-mining technique such as PSI-BLAST (Altschul *et al*, 1997) to extract relevant data from the sequence databases, to apply a neural network such as PSIPRED (McGuffin *et al*, 2000) to post-process that data, and then to use a range of statistically derived sets of parameters together with pseudo-random sampling procedures to generate potential three-dimensional models of a protein. None of these steps need to be physically relevant to the protein under study, nor do the results provide an explanation of the structures. Nevertheless, these procedures lead to extremely efficient predictions that are of great use to the biological community.

> **Scientific research itself can be thought of as a type of evolved organism…**

In fact, the development of devices to assist in the prediction of the behaviour of biological organisms—rather than to understand them—is a standard component of biological research. A significant proportion of published research has resulted from these types of engineering solution to prediction problems. We should not fool ourselves into expecting more cognitively appealing results: a deeper understanding of the workings of biological organisms might be desirable, but their evolutionary history gives us no reason to assume that we will gain any insight deeper than the level of basic chemical or physical interactions. These conclusions are tacitly assumed in the everyday research practice of biologists.

On the face of it, this might present an alarming challenge to our conception of what is good science. The accounts of the scientific method that we described earlier in this article were motivated by the belief that science has a distinctive methodology and that we are justified in believing that scientific investigation yields knowledge. But if, as we have argued, this area of biological science is not essentially in the business of giving us theories and explanations, then what grounds do we have for confidence in its results? Moreover, without a methodology, why should theology or astrology not be an acceptable form of scientific investigation?

To answer the second question first: a recognition of the distinctive ways of tackling problems in biological sciences, and in particular the ways in which they diverge from the traditional conception of science founded in physics, clearly imposes a broadening of our conception of what counts as good science. But it does not follow that 'anything goes'. What the biological sciences have essentially in common with other fields of science is their focus on predicting and manipulating the phenomena under investigation. It is therefore not the method that should control our conception of what counts as good science, but rather the measure of success. Finding a theory that explains how something works is only one way of meeting this standard. Engineering a program to predict how something behaves is another way—increasingly common in biology—of achieving the same result.

The idea that science must exhibit standards of success is not, of course, a new one (Stenger, 2006). But scientists have typically assumed that successful predictions must emerge from a theoretical understanding of the phenomena under investigation. However, a growing body of research in biology engineers mechanisms that generate predictions without such a theoretical context, and this is a significant departure both in our conception of how science should be done and in the way that it is done in other areas of science.

Yet, the lack of a unified research method should not lead us to doubt that science provides reliable knowledge. Scientific research itself can be thought of as a type of evolved organism: it is subject to various selection pressures—such as the need to produce successful results and publishable work—and is constantly contributed to and modified by the input of researchers. Crucially, under these circumstances, it regularly produces reliable results, and it is this reliability that underpins our judgement that through science we can know the world. While the environment remains competitive and science resilient, we can have confidence in the results that it yields.

REFERENCES

Alexander PA, He Y, Chen Y, Orban J, Bryan PN (2007) The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proc Natl Acad Sci USA* **104:** 11963–11968

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389–3402

Andreeva A, Murzin AG (2006) Evolution of protein fold in the presence of functional constraints. *Curr Opin Struct Biol* **16:** 399–408

Birzele F, Csaba F, Zimmer R (2008) Alternative splicing and protein structure evolution. *Nucleic Acids Res* **36:** 550–558

Csete ME, Doyle JC (2002) Reverse engineering of biological complexity. *Science* **295:** 1664–1669

Deutsch D (1998) *The Fabric of Reality*. London, UK: Penguin

Gannon F (2007) Too complex to comprehend? *EMBO Rep* **8:** 705

Goh CS, Milburn D, Gerstein M (2004) Conformational changes associated with protein–protein interactions. *Curr Opin Struct Biol* **14:** 104–109

Hartl FU, Hayer-Hartl M (2002) Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science* **295:** 1852–1858

Horrobin DF (2001) Realism in drug discovery—could Cassandra be right? *Nat Biotechnol* **19:** 1099–1100

Kellenberger E (2004) The evolution of molecular biology. *EMBO Rep* **5:** 564–569

Larranaga P *et al* (2006) Machine learning in bioinformatics. *Brief Bioinform* **7:** 86–112

Macchiarulo A, Nobeli I, Thornton JM (2004) Ligand selectivity and competition between enzymes *in silico*. *Nat Biotechnol* **22:** 1039–1045

Machamer P, Darden L, Craver CF (2000) Thinking about mechanisms. *Philos Sci* **67:** 1–25

Mazzocchi F (2008) Complexity in biology. *EMBO Rep* **9:** 10–14

McGuffin LJ, Bryson K, Jones DT (2000) The PSIPRED protein structure prediction server. *Bioinformatics* **16:** 404–405

Pearson H (2002) Surviving a knockout blow. *Nature* **415:** 8–9

Popper K (1963) *Conjectures and Refutations: The Growth of Scientific Knowledge*. London, UK: Routledge and Kegan Paul

Sarkar S (1998) *Genetics and Reductionism*. Cambridge, UK: Cambridge University Press

Schaffner K (1993) *Discovery and Explanation in Biology and Medicine*. Chicago, IL, USA: University of Chicago Press

Service RF (2007) Resurrected proteins reveal their surprising history. *Science* **317:** 884–885

Stenger VJ (2006) *The Comprehensible Cosmos: Where do the Laws of Physics Come From?* Amherst, NY, USA: Prometheus

Tompa M *et al* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* **23:** 137–144

van Regenmortel MHV (2004) Reductionism and complexity in molecular biology. *EMBO Rep* **5:** 1016–1020

Weng G, Bhalla US, Iyengar R (1999) Complexity in biological signaling systems. *Science* **284:** 92–96

Wimsatt WC (1976) Reductive explanation: a functional account. In *Re-engineering Philosophy for Limited Beings* (2007). Cambridge, MA, USA: Harvard University Press

Lawrence Kelley (left) is at the Structural Bioinformatics Group, Division of Molecular Biosciences at Imperial College London, UK.
E-mail: l.a.kelley@imperial.ac.uk
Michael Scott is at the Centre for Philosophy, School of Social Sciences, University of Manchester, UK.
E-mail: michael.scott@manchester.ac.uk

doi:10.1038/embor.2008.212
Published online 14 November 2008