Lander *BMC Biology* 2010, **8**:40
http://www.biomedcentral.com/1741-7007/8/40

BMC Biology
incorporating Journal of Biology

OPINION

The edges of understanding

Arthur D Lander*

Abstract

A culture's icons are a window onto its soul. Few would disagree that, in the culture of molecular biology that dominated much of the life sciences for the last third of the 20th century, the dominant icon was the double helix. In the present, post-modern, 'systems biology' era, however, it is, arguably, the hairball.

By hairball I refer here to those stunningly complicated network diagrams that grace the pages (and covers) of major journals with some regularity, in which the vertices or 'nodes' are annotated with symbols representing genes, proteins or metabolites, and the connectors or 'edges' are usually so numerous as to strain the resolution of monitors and printers (Figure 1).

While lacking much of the aesthetic appeal of a double helix, the hairball can be seen as iconic because it succinctly captures the distinctive flavor of systems biology. A molecular biologist and a systems biologist both construct their view of biology out of knowledge of biology's components (nodes) and knowledge of the relationships among those components (edges) (Figure 2). Where they differ is in the relative emphasis they place on each: to the molecular biologist, the answers to difficult questions are sought mainly by discovering nodes and linking them through edges that stand for qualitative causal relationships ('gene *a* turns on gene *b*'; 'enzyme *x* phosphorylates protein *y*', and so on). To the systems biologist, answers are sought mainly through the investigation of networks themselves, the behaviors of which tend to be dominated by the quantitative details of their edges more than by the physical nature of their nodes. In molecular biology, explaining the existence of a phenotype or disease by 'finding the gene(s) for it' is a plausible goal; in systems biology it is just a starting point for investigation.

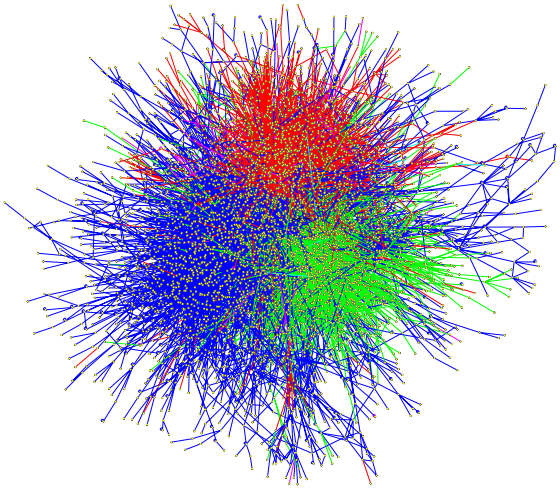

Figure 1. Human proteome, and its binding interactions. Depiction of the data as a hairball, an increasingly familiar image in the biology literature. Figure kindly provided by Nicolas Simonis and Marc Vidal, see [14].

Curiously, this distinction is often misconstrued. Among scientists, as well as the public, systems biology is frequently identified with the exploitation of high-throughput methods to gather vast amounts of data about genomes, epigenomes, transcriptomes, proteomes, metabolomes, phenomes, and the like. Sophisticated as such methodologies have become, they primarily support the tasks that molecular biologists have always faced - discovering nodes and edges. If this were all there was to systems biology, it would be hard to justify treating it as anything more than an accelerated program of molecular biology.

But there is certainly more. In driving home this point, the hairball icon is again useful, albeit with different assignments of meaning. In this interpretation, we take the nodes to represent knowledge - individual sets of data about the biological world, including facts, observations, structures, behaviors, and so on - and the edges to represent relationships, or connections, between bodies of knowledge. For example, what we know about the cell cycle and what we know about circadian behaviors such as sleep-wake cycles may be connected by virtue of the

*Correspondence: adlander@uci.edu
Center for Complex Biological Systems, Department of Developmental and Cell Biology, University of California, Irvine, CA 92697, USA

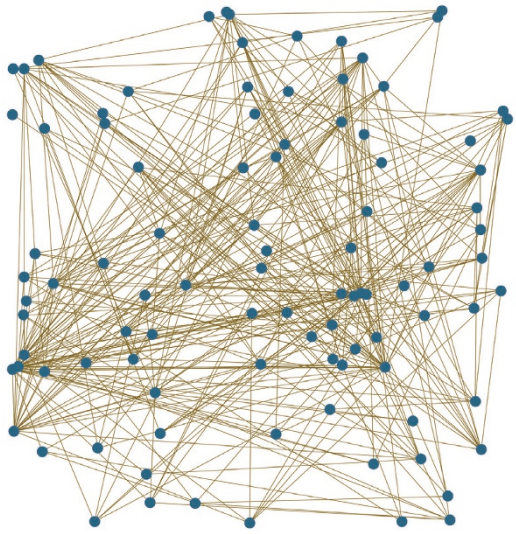

Figure 2. Hairballs are composed of nodes and edges. With nodes (blue) standing for elements such as genes, proteins, or metabolites, and edges (brown) standing for qualitative or quantitative relationships, hairballs are graphical representations of explicit models. Hairballs can also represent the act of modeling itself, with nodes that represent knowledge, and edges that represent the connections models build between bodies of knowledge.

fact that both phenomena are built upon autonomous oscillators with external inputs. These bodies of knowledge may additionally be connected to what we know about the stripes on the coats of tigers and zebras by generalizing the notion of oscillation to include oscillation in space as well as time. These can be further joined to what we know about the formation of vertebrate body segments, and the formation of patterns on seashells, through the relationship that spatial oscillations can arise through the interaction of temporal oscillation with stable spatial growth.

It is straightforward to see two things about the edges in this second kind of hairball, which serve to connect bodies of knowledge. First, they do not fundamentally stand for statements of causality (zebra stripes do not cause, nor are they caused by, shell patterns). Second, they are often peculiarly satisfying to learn about. When we appreciate that two or more very different-seeming phenomena can be treated as similar in some way, we tend to feel that we have accomplished something. This feeling has a simple name: understanding.

My purpose in offering this procedural definition of understanding is to draw an important distinction between knowledge and understanding. Factual discovery, whether in the biological sciences or any other enterprise, does not constitute understanding on its own. The student who correctly answers a question in class by downloading it off the internet with his smartphone does not necessarily understand anything.

Although I do not doubt that many of my colleagues in the sciences were lured into their professions by the thrill of discovering new knowledge, I would speculate that at least as many were attracted, as I was, by the challenge of understanding the world in new ways. It has always disappointed me that so much of the vast literature on how science is, or ought to be, practiced deals with the former goal and not the latter. Writings on the 'scientific method', whether from practicing scientists or philosophers, seem to deal mainly with how we design and perform experiments so that we can validly infer that something is or is not the case. This amounts to the question of how we arrive at potential knowledge and decide whether or not to accept it.

The question of how we create understanding out of validated bits of knowledge seems to have attracted so much less attention because, I suppose, it is easily seen as trivial. For example, if we obtain data that shutting down the activity of any of a certain set of genes blocks the ability of cells to splice pre-mRNAs, and we have previous data showing that the products of those genes physically associate in the cell to form a large supramolecular complex, we are easily drawn to view such a complex as a 'splicing machine'. Coming to this understanding is an act that does not seem to require much effort or skill. A graduate student will accomplish it as quickly as a senior professor; more quickly in some cases, because seasoned scientists tend to be more distrustful of the impulse to submerge messy facts beneath neat, orderly concepts.

There are many phrases that describe the action of replacing the messy with the simple to promote understanding: 'creating an abstraction', 'generalizing' and 'distilling a concept' come to mind, but the phrase I find most evocative is, 'building a model'. When we understand a collection of gene products as a splicing machine, we are building a model of splicing that is simpler than the underlying data set that produced it. When we understand the cell cycle as a regulated oscillator, or metabolic networks as systems for optimizing growth, we are likewise building simple models of complex processes.

Models do not arise by logical inference from data; they are acts of human creation. Any set of data can be modeled in a large (perhaps infinite) number of ways. Our reasons for choosing one over another are not to be found in the data themselves, but rather in our ideas about how a model will help us connect the data to other knowledge. This point is well illustrated in Kyle Stanford's book *Exceeding our Grasp*, [1] which investigates the origins of influential biological models that were later discarded or discredited. Stanford relates how some of the best minds in biology routinely failed to conceive of

the models that would eventually supplant their own, even when the later models would have equally well fitted all the data to which they had access.

Models are valuable in science not because they can be validated, but because they can be useful. Indeed, the entire notion of validating or invalidating models seems misguided. Models may be found inconsistent with a set of data, but that does not necessarily rob them of their utility. When we use Newton's laws of motion; when we identify protein domains as alpha helices or beta sheets; even when we refer to the concentration of a reactant in a cell, we are invoking models that only approximate reality. Yet the simplicity of these models makes them useful anyway, often more useful in day-to-day life than more complicated models that better fit the data. The idea that the best models never fit all the data was summed up 30 years ago by the statistician George Box when he said, 'all models are wrong; some are useful' [2].

My purpose in presenting this particular definition of 'model' is to contrast it with views now common among biologists, including many self-identified systems biologists (for example, [3,4]). In particular, there seems to be a prevailing view that modeling activities are dramatically accelerating in biology; that the primary use of modeling is to predict experimental outcomes that then validate or invalidate them; and that the overall goal of modeling is to generate testable hypotheses. It strikes me that all three statements are misapprehensions.

First, models have long been abundant in biology. Pick up any of the classic textbooks of molecular biology from the 1970s through the 1990s and you will typically find that half the illustrations are models of some sort or another. The difference between molecular biology and systems biology is that models in the former field are usually represented as cartoons and arrow diagrams, whereas in the latter they are more often represented as sets of equations or procedural instructions. It is not the use of modeling, *per se*, that is changing, it is the elements out of which biologists tend to build models. Such a change enables us to use our models to find different kinds of connections between bodies of knowledge. For example, with cartoon-based modeling, we can see that G-proteins and signal-controlled protein kinases are similar in that both use the thermodynamics of phosphorylation and dephosphorylation to drive irreversible switches. With equation-based modeling, however, we can see that the ultrasensitivity required for switch-like behavior can be created out of multisite phosphorylation with distributed kinetics [5].

Second, the idea that models only serve us to the extent that they make experimental predictions is, in my opinion, one of the more pernicious widely-held notions in biology today. While predictions may indeed flow from models, it is too easy for such predictions to have little bearing on a model's value. To paraphrase a bad children's joke, the observation that an amputee frog does not jump in response to verbal commands is indeed a prediction of the model that frogs hear with their legs. While the foregoing is an intentionally facetious example, it is not far off from the situation in which colleagues of mine have sometimes found themselves: forced by anonymous reviewers to make, and then test, gratuitous predictions of their models just to get their work published.

Of course, not all predictions are gratuitous. Demonstrating that a model continues to fit new data can be extremely helpful, especially when trying to choose among a range of possible models. But there are so many other ways in which models can be useful. The social scientist Joshua Epstein recently compiled a list of 16 other reasons for modeling besides prediction. Among them are to provide explanation; illuminate dynamics; suggest analogies; identify new questions; and demonstrate tradeoffs [6]. Whereas these activities have nothing to do with prediction, they have everything to do with understanding. We also learn from Epstein that preoccupation with predictive modeling is not unique to biologists. As he remarks about his colleagues, 'For some reason, the moment you posit a model, prediction - as in a crystal ball that can tell the future - is reflexively presumed to be your goal' [6].

The third misapprehension - that the ultimate goal of modeling is to generate hypotheses - is often promoted by modelers of biology themselves. According to this view, data in biology are now being gathered so rapidly, and in such a comprehensive way, that the pace at which we make experimental observations is outstripping the pace at which we usually formulate good hypotheses to test. Modeling, particularly the computational modeling of statistical correlations, is said to provide an efficient tool for finding such hypotheses amidst the mass of data [7,8]. A continual cycle of modeling, hypothesis generation, experimentation, and model refinement is proposed as the only logical way for biology to move forward.

It is true that efficient routes to hypothesis generation are greatly needed in biology. It is also true that models provide the structure within which hypotheses can be framed. Indeed, the reason a 'robot scientist' can efficiently perform impressive feats of biological hypothesis generation and testing [9] is that it is pre-programmed with basic models of how certain domains of biology work. But to characterize models solely as tools for hypothesis generation underplays the role of models as vehicles of understanding. Indeed, we could imagine a far-off future in which so much knowledge has been gathered that virtually every imaginable hypothesis has already been tested (whether intentionally or not). Would we have no need for models in such a future? To the contrary, with so much to make sense of, I would expect the need to be even greater.

How, then, should we decide when and whether to model? And if modeling is meant to forge connections between bodies of knowledge, is there a systematic way of making sure this succeeds? If the elements of one's models are only cartoons and arrow diagrams, these questions are probably fairly simple to deal with. But for models built out of sets of mathematical equations and statistical constructs, as is increasingly the case in systems biology, the answers are by no means obvious. Fortunately, they are not entirely occult either: making connections between explicit, systematic representations of complex things is the bread and butter of at least three fields outside of biology, namely mathematics, theoretical physics and 'theoretical engineering' (which includes control theory). That these theoretical disciplines have been playing an increasing role in the development of systems biology (see, for example, [10-12]) may be a sign that biology is finally ready for its own 'theory branch' [13]. This suggests that the 21st century may be remembered as a time when biology finally dedicated itself to systematic exploration, not just of the limits of knowledge, but all the way to the edges of understanding.

Published: 12 April 2010

References

1. Stanford PK: *Exceeding our Grasp: Science, History and the Problem of Unconceived Alternatives.* New York: Oxford University Press; 2006.
2. Box GEP: **Robustness in the strategy of scientific model building.** In *Robustness in Statistics.* Edited by Launer RL, Wilkinson GN. New York: Academic Press; 1979: 201-236
3. Ideker T, Winslow LR, Lauffenburger DA: **Bioengineering and systems biology.** *Ann Biomed Eng* 2006, **34:**1226-1233.
4. Aderem A: **Systems biology: its practice and challenges.** *Cell* 2005, **121:**511-513.
5. Huang CY, Ferrell JE Jr: **Ultrasensitivity in the mitogen-activated protein kinase cascade.** *Proc Natl Acad Sci U S A* 1996, **93:**10078-10083.
6. Epstein JM: **Why model?** *J Artificial Societies Social Simulation* 2008, **11:**12.
7. Kell DB, Oliver SG: **Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era.** *Bioessays* 2004, **26:**99-105.
8. Nabel GJ: **Philosophy of science. The coordinates of truth.** *Science* 2009, **326:**53-54.
9. King RD, Rowland J, Oliver SG, Young M, Aubrey W, Byrne E, Liakata M, Markham M, Pir P, Soldatova LN, Sparkes A, Whelan KE, Clare A: **The automation of science.** *Science* 2009, **324:**85-89.
10. Tomlin CJ, Axelrod JD: **Understanding biology by reverse engineering the control.** *Proc Natl Acad Sci U S A* 2005, **102:**4219-4220.
11. Mogilner A, Wollman R, Marshall WF: **Quantitative modeling in cell biology: what is it good for?** *Dev Cell* 2006, **11:**279-287.
12. Lewis J: **From signals to patterns: space, time, and mathematics in developmental biology.** *Science* 2008, **322:**399-403.
13. Wolkenhauer O, Mesarovic M, Wellstead P: **A plea for more theory in molecular biology.** *Ernst Schering Res Found Workshop* 2007**:**117-137.
14. Ferrell JE Jr: **Q&A: Systems biology.** *J Biol* 2009, **8:**2.

doi:10.1186/1741-7007-8-40
Cite this article as: Lander AD: The edges of understanding. *BMC Biology* 2010, **8:**40.