Half dozen of one, six billion of the other: What can small- and large-scale molecular systems biology learn from one another?

Ian A. Mellis[1] and Arjun Raj[2]

[1]*Perelman School of Medicine, Genomics and Computational Biology Graduate Group, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6021, USA;* [2]*Department of Bioengineering, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6321, USA*

Small-scale molecular systems biology, by which we mean the understanding of a how a few parts work together to control a particular biological process, is predicated on the assumption that cellular regulation is arranged in a circuit-like structure. Results from the omics revolution have upset this vision to varying degrees by revealing a high degree of interconnectivity, making it difficult to develop a simple, circuit-like understanding of regulatory processes. We here outline the limitations of the small-scale systems biology approach with examples from research into genetic algorithms, genetics, transcriptional network analysis, and genomics. We also discuss the difficulties associated with deriving true understanding from the analysis of large data sets and propose that the development of new, intelligent, computational tools may point to a way forward. Throughout, we intentionally oversimplify and talk about things in which we have little expertise, and it is likely that many of our arguments are wrong on one level or another. We do believe, however, that developing a true understanding via molecular systems biology will require a fundamental rethinking of our approach, and our goal is to provoke thought along these lines.

Why systems biology?

For many years now, it has been de rigueur to begin any discussion of systems biology with the question, "So, what exactly *is* systems biology?" This question surely has many answers, but perhaps a more useful question might be, "Why do we need systems biology?" …or, more generally, "Why do we need anything new?" After all, the now-standard approaches from molecular biology have provided us with unprecedented knowledge of the inner workings of the cell, transforming our understanding of biology along the way. Armed with the tools of biochemistry, the scientists in the vanguard of molecular biology's golden era worked out the structure of DNA, the genetic code, how DNA is replicated, how genes express, how cells move, and countless other fundamental pieces of the machinery that makes cells work. Importantly, these discoveries enabled precise manipulations—using the genetic code, for instance, we can use our understanding of the cell's machinery to make our own proteins.

With the development of these tools came the potential to carry out studies of ever greater scope (Snyder 2013). The paradigm here is the scientific story that connects, say, a mutation in a particular gene to an organismal phenotype via a biochemical mechanism. A shining example is that of cystic fibrosis, a heritable disease in which mutations to the *CFTR* gene, a cAMP-activated chloride channel, cause mucus to become sticky (among other effects), thus causing the human disease condition. Through careful molecular biology and biochemistry, scientists were able to delineate a clear path from mutation to biochemical defect to disease (Guggino and Stanton 2006). As the field matured, however, such examples became rarer and more tenuous; and there was a growing realization that further progress would require an understanding not just of biological mechanism, but also biological regulation. For example, although our understanding of the mechanics of transcription was fairly solid, we instead needed to know why one gene was transcribed while another one was not. In this context, the goal is to understand the regulatory interactions between molecules; and for us, it is here that the conceptual transition from molecular biology to molecular systems biology begins.

One approach to understanding biological regulation, which we call "small-scale" systems biology, came into vogue in the late 1990s. This approach, characterized by a heady blend of experimental observation and mathematical modeling, was predicated on the notion that a complete understanding of regulation would require quantitative models. Drawing analogies to electrical engineering, the object of study was the biological regulatory circuit, and the goal was to understand the principles underlying the organization of these circuits. This approach has yielded many celebrated successes ranging from bacterial chemotaxis to metazoan development. Some of our favorites include the use of integral feedback control to regulate tumble frequency in *E. coli* (Barkai and Leibler 1997; Alon et al. 1999), precision and scaling in development (Gregor et al. 2007a,b; Ben-Zvi et al. 2008; Little et al. 2013), design principles of cell signaling pathways (Cai et al. 2008; Mettetal et al. 2008; Muzzey et al. 2009; Sprinzak et al. 2010), cell-cycle regulation (Doncic and Skotheim 2013; Doncic et al. 2015), and basic mechanisms of cellular growth, metabolism, and homeostasis (Youk and van Oudenaarden 2009; You et al. 2013; Campos et al. 2014; Soifer et al. 2014; Padovan-Merhar

Corresponding author: arjunraj@seas.upenn.edu
Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.190579.115. Freely available online through the *Genome Research* Open Access option.

25:1466–1472 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/15; www.genome.org

et al. 2015). We also have a predilection for single-cell biology (Raj and van Oudenaarden 2008; Balázsi et al. 2011).

Beautiful as these stories may be, how do they differ from what researchers have been doing for years? Do they really represent the first steps toward a new hybrid field, or are they just standard molecular biology with some equations and error bars for decoration? The latter is a fair charge in many instances, but to us, the conceptual difference is that by building a quantitative understanding of biological regulation, we can build a rational, quantitative model of the cell, i.e., we can build the whole from the sum of the parts.

So how close are we to this grand ideal? Here, we will argue that the advent of "large-scale" systems biology—namely, various forms of omics-level analyses—has helped cast doubt upon the very idea that we can assemble these simple models into a global picture of biological regulation, particularly in metazoans. Yet at the same time, we believe that large-scale systems biology has not yielded any viable alternatives, but has instead just cataloged increasing complexity rather than reduced it. In this perspective, we will expand on these arguments with selected examples from genetic algorithms, genetics, transcriptional network analysis, and genomics. We will conclude with proposals for new directions for molecular systems biology that eclipse our natural mental capacities, perhaps guided by artificial intelligence–aided interpretation of data that may contain structure.

An inconvenient truth illustrated in silicon

A central tenet of systems biology is the idea that regulation in molecular biology is modular, meaning that individual components may operate independently of one another and can be strung together in a rational way to produce higher-level functions. In this context, it is easy to see why the integrated electronic circuit —a triumph of intelligent modular design—has served as a natural conceptual framework for small-scale systems biology. They are exceedingly complex regulatory devices; yet this complexity arises from the composition of smaller, comprehensible modular components integrated according to a set of design principles. The hope is that we can gain the same detailed level of understanding of biological regulatory circuits as we have with electrical circuits by isolating and understanding regulatory modules and their interconnections.

Yet a quick look at the world around us reveals the enormous differences between human-designed objects and those designed by evolution in the natural world. Given these differences, is there any reason we might expect to see similarly modular regulatory behavior in the muddled molecular soup of the cell, shaped by the random forces of evolution rather than a rational agent?

This is in general a difficult question to answer, because we typically look at evolution in the biological context and human design in the context of man-made objects like electrical circuits. But what would happen if you designed electrical circuits by evolution? Would they resemble human design?

Indeed, would we even expect an evolutionary process to yield modular circuits in silicon? In a very interesting set of experiments from the latter half of the 1990s, Adrian Thompson examined exactly that by attempting to evolve an electronic circuit (Thompson 1997). Specifically, the goal was to evolve a circuit capable of discriminating electronic "tones" (essentially, a low frequency and high frequency signal), and the evolutionary substrate for his experiments was the field programmable gate array (FPGA), a chip that is itself reprogrammable through software. The idea was to start with a randomly "programmed" chip and then evolve the features on the chip using a genetic algorithm to see if it would eventually be able to perform the tone-discrimination task (Fig. 1A). Amazingly enough, in about 4100 generations, the chip had evolved to the point at which it could perform the task with very high fidelity, which was particularly impressive given the small number of potential circuit elements it was given to work with (i.e., 100). (In fact, Thompson noted that many thought it would be impossible for such a small circuit to perform this task.) How did the circuit achieve this task? It was unfortunately rather hard to say exactly how, but one quick conclusion is that the solution was neither modular nor readily comprehensible through standard digital circuit design principles, with strange waveforms appearing throughout the evolutionary process. Thus, it was the first strike against modularity.

The beautiful thing about evolving a circuit in silicon is that it enabled Thompson (1997) to dissect the evolutionary process in ways that would be very difficult to do in the messy world of molecular biology. For instance, he could perform the equivalent of looking for "phenotypes" of "knock outs" with rapidity and precision, and here things got even more interesting. One of the first things Thompson did was prune the network by clamping the output of all circuit elements that did not affect the performance of the circuit. Surprisingly, this minimal network included five elements that were not physically connected to the circuit at all, at least not in the conventional sense (Fig. 1B). Most of these elements had relatively small but quantifiable effects on circuit performance; one of them had a very large effect, even though its output had no connection to the circuit whatsoever. The interpretation is that evolution took advantage of the underlying physics (which we typically ignore in modular design) to arrive at solutions that appear incomprehensible and foreign, even to circuit engineers. Indeed, moving the exact same circuit to a different

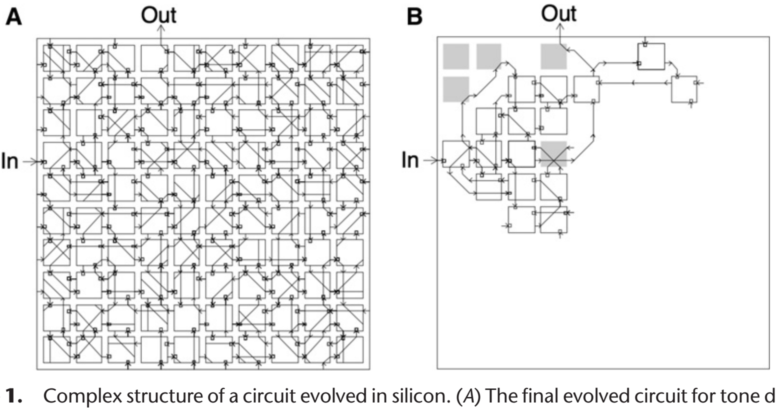

Figure 1. Complex structure of a circuit evolved in silicon. (*A*) The final evolved circuit for tone discrimination, a 10 × 10 array of cells. All connections that link an input and an output are marked. (Arrow from a cell) connection driven by origin cell's function; (square at arrowhead) connection selected as input for recipient cell function. (*B*) Minimal necessary components of final evolved circuit for tone discrimination. Missing cells can be fixed at constant value without affecting performance; gray cells cannot be fixed at constant value without affecting performance, although the cell has no path connecting to marked minimal necessary functional cells (Thompson 1997).

part of the chip resulted in poorer performance that could be improved by a bit of further evolution, a feature strikingly reminiscent of approaches in synthetic biology (Dougherty and Arnold 2009). Still, although Thompson (1997) should of course not be taken as a literal model for evolution per se, it is worth noting that it produced a strikingly organic result, and one that appears to also hold on to its secrets in much the same way.

Limitations of the small-scale systems biology approach

What are the implications of Thompson's findings for small-scale systems biology (Thompson 1997)? We believe the primary lesson is that the notion that evolution favors easily discernible and well-isolated regulatory modules may be fundamentally wrong. Although we might be able to make some sense of a small subset of the regulatory network, we believe that in most cases in metazoans, small-scale systems biology has not led to an understanding sufficiently detailed to allow one to clearly define regulatory modules or how they might interact, much like Thompson's evolved circuits.

Take, for example, the standard formula for a small-scale systems biology paper: Start with an interesting phenomenon and make some quantitative measurements. Develop a mechanistic model, often mathematical, to explain the data. Mine this model for a prediction. Make a perturbation to test this prediction and verify it experimentally. Although these stories are often very appealing, when examined in detail, one often finds that perturbations seldom yield complete and definitive results, hence leaving us with several holes in the story. These holes typically remain unfilled, and so it is difficult to know to what extent the regulatory "module" identified in the story is truly isolated from other putative modules. Indeed, these holes call into question the very notion that modules exist at all.

Take an example from our own work (Raj et al. 2010), which we focus on for the purposes of self-deprecation rather than self-promotion. We tried to explain variability in the gut development pathway in *C. elegans* in organisms harboring a mutated version of a particular transcription factor. This phenotypic variability, known as incomplete penetrance, is a common fact of life in genetics, with many, if not most, regulatory mutants showing such partial effects. We made quantitative measurements of transcription factor expression in a series of mutants, showing that variability in expression of a downstream regulator, subjected to a threshold, led to the decision of whether or not to form gut cells in the mutant embryo through the expression of the "gut master regulator" *elt-2*. A prediction of this model was that reducing the variability would result in more embryos surpassing the threshold. Sure enough, by further removing an inhibitor of the downstream regulator, we were able to reduce the variability, pushing more embryos above the threshold and leading to more gut cell formation. All in all, it was a fairly standard exercise in molecular systems biology. Yet, are we any closer to understanding variability in gut cell development? We would argue not. We know a few factors that seem to both incur this variability and then potentially manipulate the variability, but none of these perturbations give us anywhere close to a complete understanding of what governs this variability—nothing we did could restore wild-type precision fully, nor were the effects limited in scope.

Scientists often attribute these untidy results to the relative imprecision of our experimental tools (Lazebnik 2002). Perhaps that is true in some cases, but we feel the evidence points to messiness being a fundamental feature of biological regulation. In the case of gut development, decades of painstaking genetics (Maduro and Rothman 2002; Maduro 2006) provided us with a relatively simple regulatory pathway that served as the framework for our results. Yet even these genetic foundations still contain "mystery factors" that we know must exist (and probably several more that we do not know exist), but have not yet identified after almost 20 years (Maduro and Rothman 2002).

Indeed, upon closer inspection, these simple pathways often reveal layers of deep complexity and interconnectedness that are at odds with the notion of modularity. One concept that is most associated with modularity is that of the "master regulator," in this case, a transcription factor whose expression is necessary and sufficient for a particular phenotype of interest. In the case of the gut, early experiments revealed *elt-2* to be a candidate for the master regulator of gut formation (Fukushige et al. 1998). Surprisingly, it then turned out that the *elt-2* knockout worm still expressed some downstream factors and had a reasonable approximation of a nascent gut (Fukushige et al. 1998; Sommermann et al. 2010). Attention then turned to the role of a the seemingly redundant transcription factor *elt-7*, with the double knockout of *elt-2* and *elt-7* showing a far more profound lack of gut phenotype, except for the well-differentiated gut cells interspersed between the cells displaying the mutant phenotype (Sommermann et al. 2010). Perhaps then there is also a role for *elt-4* (Maduro and Rothman 2002)? One can get into semantic discussions over necessity and sufficiency and the definition of master regulators (Chan et al. 2013), but we think this example nicely illustrates the fact that redundancy and partial effects are more the rule than the exception with respect to so-called master regulators. Indeed, in many cases, close inspection of the details reveals that many other examples of master regulators are somewhat less clean and simple than previously believed.

Some may counter our arguments by pointing out that we are at the very beginning of this field, developing the initial knowledge of the dominant parts and players that we will then refine and add to over time toward a more complete understanding (Brenner 2010). Rob Phillips is a strong proponent of this line of thought, pointing out that many fields of study look messy initially until decades of hard, careful work brings a systematic order to them. Impressively, his group has shown that careful analysis of transcriptional factor concentration could explain transcriptional regulation in *E. coli* through a single governing principle (Brewster et al. 2012, 2014). It is also possible that currently mysterious regulatory behavior may have explanations involving other forms of biochemical and physical interactions than typically considered (Frechin et al. 2015).

Are these the first steps along the path to a complete explanation of transcriptional regulation? Is "complexity" just a word we trot out whenever we refuse to think harder about the problem? Or are these successes one-offs or limited in scope to simpler prokaryotic systems and utterly useless in the face of metazoan complexity? We think it is hard to say at this point. An oft-repeated truism from George E.P. Box is that all models are wrong, but some are useful. We think the difficulty lies in the definition of useful (Box and Draper 1987). As a means to roughly explain some effects in a particular regulatory system, our current models are useful. As a building block for a larger model to truly explain, for instance, how a developmental gene network attains such high levels of precision, our current models are still largely useless. We wonder whether such higher-order models will ever emerge from the paradigm of combining modular building blocks because biological regulation may be intrinsically nonmodular and thus perhaps not understandable by the framework used by small-scale systems biology.

Genomics and the revelation of dense interconnectedness

The arrival of the genomic era has in many ways laid these facts bare. Take the example of differential gene expression analysis. Now that we have the ability to accurately measure differences in gene expression across the genome, it is clear that the consequences of virtually any perturbation are seldom relegated to one or a few genes, but rather spread across large numbers of genes, often numbering in the thousands. Moreover, these sets of genes almost never fall into clear mechanistic subgroups, but rather only show “enrichment” for various cellular functions that typically overlap significantly with other perturbations.

It is certainly possible that the majority of these differences are completely inconsequential, as nicely argued by Atay and Skotheim (2014). In this view, there is a core set of circuits underlying cellular regulation surrounded by a bunch of irrelevant noise. We are not so sure, however, that this view is completely supported by the evidence.

We offer another recent example, CellNet, a computational framework developed for using genome-wide expression profiles analysis to help understand and manipulate cell types (Cahan et al. 2014; Morris et al. 2014). CellNet takes as input a large number of gene expression profiles spanning several different cell types and several different experimental conditions. Using these data, it attempts to construct a gene regulatory network associated with each cell type using expression variation to infer regulatory links. The authors applied CellNet in two ways, both of which we believe argue against simple models of gene regulation. First, using CellNet, they show that interconverting cells from one type to another by expressing particular master transcription factors (in this case, fibroblasts to neurons via ectopic expression of *Ascl1*, *Nr4a2*, and *Lmx1a*) led to cells that still had traces of the fibroblast gene expression program. Direct differentiation from embryonic stem cells showed no such defects. This shows that activation of the network by these transcription factors was incomplete, and thus that differentiation depends on more than just the expression of a few major players. Second, in the case of B cell to macrophage conversion, they showed that CellNet could generate a list of candidate interventions to enhance conversion. Experiments showed that performing those interventions worked as predicted; it seems reasonable to assume that the more interventions one could perform simultaneously, the better the results. Again, these results suggest that properties such as cell fate depend on the values of many cellular parameters, and further, that precise manipulation of those properties may require control over all of those parameters.

We think that the bias toward isolating single dominant factors stems from an inherent desire to develop scientific stories, which are invariably more satisfying when they have a single or few protagonists. Experimental geneticists typically model variants leading to big phenotypes with high penetrance, ignoring or perhaps not even detecting variants of lesser or partial effect. Molecular biologists apply many of the same experimental approaches that were well suited to working out the basic machinery of the cell but may be less well suited to understanding regulation. A particularly stark example of the limitations of this approach is in the mechanistic study of cancer, which has led to an incredible accumulation of knowledge about the molecular basis for regulating cellular processes such as proliferation, death, and disease (Hanahan and Weinberg 2000). Yet, with all this knowledge, we are still unable to cure most actual clinical human cancers, and there is a growing appreciation that detailed mechanistic models have largely failed to capture the full complexity of the disease (Weinberg 2014). Because of these biases, both scientific and methodological, it is still unclear how many forms of biological regulation are built from the sum of many smaller effects.

Unbiased—and often indecipherable

Genomics has also allowed us to remove some of these biases, perhaps most notably through the use of genome-wide association studies of quantitative traits, such as height or blood cholesterol levels. Here, the goal is to start with the quantitative trait, then look for the genetic variants underpinning its variation, many of which are in regulatory noncoding DNA. The results of genome-wide association studies have left us with the impression that the major players we seek in molecular biology exist but are rare, with most studies failing to find large-effect-size variants for most common quantitative traits. This has left us with just a few mechanistic crumbs and the general feeling that most traits are indeed composed of large numbers of R.A. Fisher’s variants of small effect, as though phenotypes are composed largely from the little gray squares of Thompson’s evolved circuits (Fig. 1B; Fisher 1930; Thompson 1997).

Perhaps the most classic example is that of human height. Height has a strongly genetic basis, with a “narrow-sense” heritability of around 0.80 (Silventoinen et al. 2003; Visscher et al. 2006). The approach of the molecular biologist would be to then look for mutants that are abnormally large or small, which would identify pathways associated with gigantism or dwarfism. Yet these results would yield little understanding of the heritability of more common variation in height, which genome-wide association studies have the potential to reveal. However, recent genome-wide association studies show (e.g., Lango Allen et al. 2010) that even the combined effects of 180 identified variants only explain ∼12.5% of the genetic differences in human height. This is not to say that the genetic basis of height is magical: If one includes all possible variants, one can explain a large fraction of the heritability (Yang et al. 2010). Rather, this points to height being a composition of a very large number of very small effects, and the same story has come up in analyses of many other traits. And what of the molecular basis of normal variation in human height? Several experiments done both before the advent of genome-wide association studies and afterward as follow-ups on identified loci have suggested that several of the SNPs identified in these studies have functional effects in pathways that can plausibly be linked to height, such as mitosis, mesoderm and skeletal development, and a plethora of signaling pathways, including those controlled by various growth factors, among others (Wood et al. 2014). Still, there is no single pathway to point to that provides a simple story explaining height, and as such, no simple therapeutic intervention to enable us to manipulate height.

This is not to say that genome-wide association studies have not revealed variants of immediate biomedical interest. There are many examples to choose from, including Musunuru and colleagues’ work taking a hit from a genome-wide association study, showing that the SNP changed expression of a particular gene, which then altered lipid levels in the blood (Musunuru et al. 2010). This example and others like it provides a beautiful arc from discovery to mechanism and is in many ways an ideal that the field aspires to. Yet, this is much more the exception than the rule, with perhaps most genetic variants having a spectrum of effect strengths.

There are of course many debates as to exactly why the field of quantitative genetics is filled with more of a murky haze than a set of smoking guns, and we leave it to those more qualified than us to continue those debates. (We do wonder if these very complex mappings from genotype to phenotype may actually reflect the advantages of a distributed manner of encoding.) However, from our relative outsiders' perspective, we believe these findings fit well with our general thesis that biological regulation is far less story-like than we would like it to be, and there is a distinct possibility that many if not most regulatory systems have almost no dominant, easily rationalizable stories to be found at all.

Now that we know the unknown unknowns, what do we really know?

Should we then dispense with the very notion of small-scale systems biology and plunge headlong into a data-first future? Ultimately, we think this depends on the nature of the question at hand and the type of understanding we hope to derive. At one end of the spectrum is a purely operational level of understanding, one in which we learn just enough to be able to manipulate cells in ways we find useful or amusing—something that may require less in the way of deep understanding. On the other end of the spectrum is the search for universal laws and design principles along the lines of Newton's laws of motion. How close are we to the latter?

We think many of us got into basic science to pursue fundamental truths, and it was not uncommon for a time to hear the claim that we are on the cusp of a Newtonian revolution in biology. Perhaps. Certainly, as a purely theoretical matter, if we were able to measure absolutely everything about a cell, the laws of chemistry would likely enable us to produce a completely predictive model of cellular function, and there are promising attempts at simulating relatively simple organisms such as bacteria (Karr et al. 2012). Yet, the development of a complete model incorporating all the complexities of a metazoan cell seems very distant at this point, and so our search for fundamental truths must be for some simplified effective representation. It is, of course, an open question as to whether universal truths such as Newton's laws even exist for cellular regulation—and if they do exist, whether we will be able to understand them. We believe our earlier arguments further support the premise that systems engineered through evolution need not be modular nor follow well-defined design principles. It is true that computational studies have shown that evolution can potentially favor modular solutions (Variano and Lipson 2004; Kashtan and Alon 2005; Clune et al. 2013), but we wonder whether the constraints imposed by models cannot reflect the ability of natural systems (or even Thompson's circuits) to take advantage of complex underlying chemistry and physics. Either way, whether one wishes to find a few greater truths or a passel of smaller ones, we find we are in a state where we suffer not from a paucity of data, but from a paucity of frameworks—theories, really—by which to understand those data.

Currently, most of our approaches to dealing with large amounts of data essentially boil down to statistical methods for extracting associations, often using increasingly sophisticated techniques from machine learning to try and generate hypotheses and insights. Yet, as Gautham Nair, a former postdoc in our laboratory, once quipped, "Would Newton have discovered the theory of gravity through machine learning?" As a related question, does the theory of gravity have a *P*-value? It is perhaps instructive to look at another example from physics: the Large Hadron Collider's search for the Higgs boson. The Large Hadron Collider produces data at an almost unfathomable rate, and yet *the vast majority of it is discarded and deemed irrelevant*. This is because the theoretical foundations of the experiment are so strong that we are able to parse this data down to the specific events that are most relevant to proving, for instance, that the Higgs boson exists. Of course, processing this data requires extremely sophisticated statistical treatment of the data, but that is more a matter of analysis than the derivation of scientific truth. We think it very unlikely that one would be able to derive all of particle physics just by drinking directly from the fire hose of particle collider data. Closer to home, one of our favorite examples of small-scale systems biology is Cai and colleagues' lovely result showing that frequency modulation of bursts of nuclear localization can coordinate expression across a very large number of genes (Cai et al. 2008). It seems similarly unlikely that one could arrive at this result simply by combing through reams of high-throughput expression data.

Our analogies have flaws, but we find most data-first counterarguments are unsatisfying. One might argue with our point about, for example, Newton's theory of gravitation, saying that it would have been impossible to even conceive of the theory without the huge collection of data on the movement of heavenly bodies. It is true that the data were there first, but it is unclear that all those data were *required* for the conception of the theory, or rather served as post hoc confirmation. In this case, all the data required would be those showing a discrepancy with the current model for planetary motion; similarly, the discovery of alternative splicing did not require deep RNA-sequencing. Another argument often cited as a benefit of data-first approaches is that they are not biased in favor of any particular outcome. Although we agree that the genomics tools applied in data-first approaches are extremely powerful tools for discovery (much as are genetic screens and biochemical purifications), we believe that an approach not directed toward any particular scientific question is unlikely to provide any conclusive answers (Brenner 2010; Weinberg 2010; Graur et al. 2013). (Although genomics-style research is perhaps most often criticized for data-first approaches, many other areas of biomedical research suffer the same issues, but are perhaps less well known or glamorous, thus attracting less controversy.) We think this underscores our belief that no matter what the technical approach, strong experimental design with a question in mind is still a requirement.

Rise of the machines?

With all of that said, largely statistical modes of analysis that dominate the analysis of large data sets these days are an easy target for scorn until we are faced with the challenge of actually analyzing said data ourselves. Why has deriving insight from data proven so challenging? Is it perhaps the limitations inherent to our own human brains? For instance, most human brains have limited capacity to reason beyond two (or sometimes three) dimensions. Indeed, it is for this reason that researchers have invested much effort into developing two-dimensional visualization techniques of high dimensional data like t-SNE (Van der Maaten and Hinton 2008) with applications in biology (Amir et al. 2013) in the hopes that our brains' capacity for deriving insights from 2D presentations may somehow reveal something. However, just as taxonomy is not biology, so too classification is not understanding; and it is important to separate the visualization of data with our quest to understand it. Therein lies the challenge: There is no reason to believe that the biology of, say, gene regulation is inherently understandable in some 2D manifestation. Perhaps, however, there is hope that computation may develop to the point at which computers can actually help us develop insights directly from data. Lest this sound like a Pollyannaish vision of the future, it is worth mentioning that Hod Lipson's group has demonstrated the ability to algorithmically derive mathematical descriptions of physical laws—including, Newton's 2nd law (!)—directly from motion tracking of pendulums and other such devices (Schmidt and Lipson 2009). (It is a delightful irony that the group used genetic algorithms to make these discoveries.) Applications to biology may yield new biological laws we may never have envisioned otherwise (Schmidt et al. 2011). Or perhaps we may draw inspiration from advances in computer vision, in which very large data sets coupled with large neural networks have led to stunning advances in the ability of computers to parse natural images (Deng et al. 2009; Russakovsky et al. 2014), with these programs now able to identify objects in images with startling accuracy. Recent iterations are in fact also able to parse semantics from those images. Of course, such "narrow" artificial intelligence often still pales in comparison to the power of the adult human brain in general (although in some instances can outperform even the best human). However, computational architectures are also free from the constraints that our physiology imposes and may be able to "see" patterns in higher dimensions that we simply cannot intuit without help. CellNet (Cahan et al. 2014; Morris et al. 2014) and other network frameworks (Carter et al. 2013; Carvunis and Ideker 2014) may portend the arrival of such aids to intuition. Such methods are still in their infancy, but we believe they may ultimately provide the tools required to help us derive meaning from the highly multidimensional data that is increasingly ubiquitous in molecular biology. Whatever the approach may ultimately be, we believe that the complete reverse engineering of regulation in molecular biology will require fundamentally new computational aids that enable us to extract some order from the seemingly endless complexity we are now faced with.

We also think that synthetic biology has the potential to inform our understanding. Currently, some synthetic reconstructions of biology are able to capture aspects of real biology to some degree, allowing us to test biological hypotheses in a rigorous fashion. It also may be that the incorporation of new, computer-based insights can harness complexity to yield a far greater degree of control over biological systems than is currently possible. This may reveal, however, the requirement of new forms of manipulations that enable us to produce complex, multifactorial perturbations.

Is there any hope left for small-scale systems biology?

What, then, to make of small-scale systems biology? Is it worth our continued pursuit? The answer for us is yes. We do think, though, that small-scale systems biology will look a bit different in the future. Currently, the most visible differences between small-scale systems biology and large-scale systems biology have been methodological, with a clear dividing line between fluorescent protein reporters, single molecule readouts, and tinkering with the genetics of model organisms on the one hand and large-scale consortium-driven omics approaches on the other. Yet these are just differences in technology, and the differences in style are perhaps driven more by the "design by committee" approach required for what used to be very expensive large-scale experiments. As omics technologies become cheaper, this gap is shrinking, and large-scale data is becoming more accessible to the do-it-yourself style more typically associated with small-scale systems biology. At the same time, developing new quantitative frameworks to understand what these data are telling us will still be critical, and we think it is important to keep an open mind as to what those frameworks may look like. We still believe in the goal of making quantitative models to reveal principles of cellular behavior; and perhaps through the incorporation of omics technology and new computational techniques to augment our intuition, we will be able to synthesize our small-scale models into a more complete picture. Of course, it is also possible that we may never be able to scale and integrate our models. Perhaps this is okay. Science is also about appreciating the beauty of solving puzzles, be they large or small.

Acknowledgments

We thank Jan Skotheim and Rob Phillips for discussions that informed the writing of this manuscript, and Uschi Symmons, Chris Hsiung, Paul Ginart, and other members of the Raj Laboratory for critical feedback. I.A.M. acknowledges support from a National Institutes of Health (NIH) MSTP T32 GM-07170 training grant, and A.R. acknowledges support from an NIH New Innovator grant 1DP2OD008514, a National Science Foundation NSF CAREER award, and an R33 EB019767 (NIH–NIBIB).

References

Alon U, Surette MG, Barkai N, Leibler S. 1999. Robustness in bacterial chemotaxis. *Nature* **397:** 168–171.

Amir el-AD, Davis KL, Tadmor MD, Simonds EF, Levine JH, Bendall SC, Shenfeld DK, Krishnaswamy S, Nolan GP, Pe'er D. 2013. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat Biotechnol* **31:** 545–552.

Atay O, Skotheim JM. 2014. Modularity and predictability in cell signaling and decision making. *Mol Biol Cell* **25:** 3445–3450.

Balázsi G, van Oudenaarden A, Collins JJ. 2011. Cellular decision making and biological noise: from microbes to mammals. *Cell* **144:** 910–925.

Barkai N, Leibler S. 1997. Robustness in simple biochemical networks. *Nature* **387:** 913–917.

Ben-Zvi D, Shilo BZ, Fainsod A, Barkai N. 2008. Scaling of the BMP activation gradient in *Xenopus* embryos. *Nature* **453:** 1205–1211.

Box GEP, Draper NR. 1987. *Empirical model building and response surfaces*, p. 424. John Wiley & Sons, New York.

Brenner S. 2010. Sequences and consequences. *Philos Trans R Soc Lond B Biol Sci* **365:** 207–212.

Brewster RC, Jones DL, Phillips R. 2012. Tuning promoter strength through RNA polymerase binding site design in *Escherichia coli*. *PLoS Comput Biol* **8:** e1002811.

Brewster RC, Weinert FM, Garcia HG, Song D, Rydenfelt M, Phillips R. 2014. The transcription factor titration effect dictates level of gene expression. *Cell* **156:** 1312–1323.

Cahan P, Li H, Morris SA, Lummertz da Rocha E, Daley GQ, Collins JJ. 2014. CellNet: network biology applied to stem cell engineering. *Cell* **158:** 903–915.

Cai L, Dalal CK, Elowitz MB. 2008. Frequency-modulated nuclear localization bursts coordinate gene regulation. *Nature* **455:** 485–490.

Campos M, Surovtsev IV, Kato S, Paintdakhi A, Beltran B, Ebmeier SE, Jacobs-Wagner C. 2014. A constant size extension drives bacterial cell size homeostasis. *Cell* **159:** 1433–1446.

Carter H, Hofree M, Ideker T. 2013. Genotype to phenotype via network analysis. *Curr Opin Genet Dev* **23:** 611–621.

Carvunis AR, Ideker T. 2014. Siri of the cell: what biology could learn from the iPhone. *Cell* **157:** 534–538.

Chan SS, Kyba M. 2013. What is a master regulator? *J Stem Cell Res Ther* **3:** 114.

Clune J, Mouret JB, Lipson H. 2013. The evolutionary origins of modularity. *Proc Biol Sci* **280:** 20122863.

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. 2009. ImageNet: a large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, pp. 248–255. IEEE, Miami, FL.

Doncic A, Skotheim JM. 2013. Feedforward regulation ensures stability and rapid reversibility of a cellular state. *Mol Cell* **50:** 856–868.

Doncic A, Atay O, Valk E, Grande A, Bush A, Vasen G, Colman-Lerner A, Loog M, Skotheim JM. 2015. Compartmentalization of a bistable switch enables memory to cross a feedback-driven transition. *Cell* **160:** 1182–1195.

Dougherty MJ, Arnold FH. 2009. Directed evolution: new parts and optimized function. *Curr Opin Biotechnol* **20:** 486–491.

Fisher RA. 1930. *The genetical theory of natural selection*. Clarendon Press, Oxford, UK.

Frechin M, Stoeger T, Daetwyler S, Gehin C, Battich N, Damm E-M, Stergiou L, Riezman H, Pelkmans L. 2015. Cell-intrinsic adaptation of lipid composition to local crowding drives social behaviour. *Nature* **523:** 88–91.

Fukushige T, Hawkins MG, McGhee JD. 1998. The GATA-factor *elt-2* is essential for formation of the *Caenorhabditis elegans* intestine. *Dev Biol* **198:** 286–302.

Graur D, Zheng Y, Price N, Azevedo RBR, Zufall RA, Elhaik E. 2013. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol* **5:** 578–590.

Gregor T, Tank DW, Wieschaus EF, Bialek W. 2007a. Probing the limits to positional information. *Cell* **130:** 153–164.

Gregor T, Wieschaus EF, McGregor AP, Bialek W, Tank DW. 2007b. Stability and nuclear dynamics of the bicoid morphogen gradient. *Cell* **130:** 141–152.

Guggino WB, Stanton BA. 2006. New insights into cystic fibrosis: molecular switches that regulate CFTR. *Nat Rev Mol Cell Biol* **7:** 426–436.

Hanahan D, Weinberg RA. 2000. The hallmarks of cancer. *Cell* **100:** 57–70.

Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B, Assad-Garcia N, Glass JI, Covert MW. 2012. A whole-cell computational model predicts phenotype from genotype. *Cell* **150:** 389–401.

Kashtan N, Alon U. 2005. Spontaneous evolution of modularity and network motifs. *Proc Natl Acad Sci* **102:** 13773–13778.

Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, et al. 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467:** 832–838.

Lazebnik Y. 2002. Can a biologist fix a radio?—Or, what I learned while studying apoptosis. *Cancer Cell* **2:** 179–182.

Little SC, Tikhonov M, Gregor T. 2013. Precise developmental gene expression arises from globally stochastic transcriptional activity. *Cell* **154:** 789–800.

Maduro MF. 2006. Endomesoderm specification in *Caenorhabditis elegans* and other nematodes. *Bioessays* **28:** 1010–1022.

Maduro MF, Rothman JH. 2002. Making worm guts: the gene regulatory network of the *Caenorhabditis elegans* endoderm. *Dev Biol* **246:** 68–85.

Mettetal JT, Muzzey D, Gómez-Uribe C, van Oudenaarden A. 2008. The frequency dependence of osmo-adaptation in *Saccharomyces cerevisiae*. *Science* **319:** 482–484.

Morris SA, Cahan P, Li H, Zhao AM, San Roman AK, Shivdasani RA, Collins JJ, Daley GQ. 2014. Dissecting engineered cell types and enhancing cell fate conversion via CellNet. *Cell* **158:** 889–902.

Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, Li X, Li H, Kuperwasser N, Ruda VM, et al. 2010. From noncoding variant to phenotype via *SORT1* at the 1p13 cholesterol locus. *Nature* **466:** 714–719.

Muzzey D, Gómez-Uribe CA, Mettetal JT, van Oudenaarden A. 2009. A systems-level analysis of perfect adaptation in yeast osmoregulation. *Cell* **138:** 160–171.

Padovan-Merhar O, Nair GP, Biaesch AG, Mayer A, Scarfone S, Foley SW, Wu AR, Churchman LS, Singh A, Raj A. 2015. Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. *Mol Cell* **58:** 339–352.

Raj A, van Oudenaarden A. 2008. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135:** 216–226.

Raj A, Rifkin SA, Andersen E, van Oudenaarden A. 2010. Variability in gene expression underlies incomplete penetrance. *Nature* **463:** 913–918.

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, et al. 2014. ImageNet large scale visual recognition challenge. arXiv. **cs.CV**.

Schmidt M, Lipson H. 2009. Distilling free-form natural laws from experimental data. *Science* **324:** 81–85.

Schmidt MD, Vallabhajosyula RR, Jenkins JW, Hood JE, Soni AS, Wikswo JP, Lipson H. 2011. Automated refinement and inference of analytical models for metabolic networks. *Phys Biol* **8:** 055011.

Silventoinen K, Sammalisto S, Perola M, Boomsma DI, Cornes BK, Davis C, Dunkel L, De Lange M, Harris JR, Hjelmborg JVB, et al. 2003. Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Res* **6:** 399–408.

Snyder SH. 2013. Science interminable: Blame Ben? *Proc Natl Acad Sci* **110:** 2428–2429.

Soifer I, Robert L, Barkai N, Amir A. 2014. Single-cell analysis of growth in budding yeast and bacteria reveals a common size regulation strategy. arXiv. **q-bio.CB**.

Sommermann EM, Strohmaier KR, Maduro MF, Rothman JH. 2010. Endoderm development in *Caenorhabditis elegans*: The synergistic action of ELT-2 and -7 mediates the specification→differentiation transition. *Dev Biol* **347:** 154–166.

Sprinzak D, Lakhanpal A, Lebon L, Santat LA, Fontes ME, Anderson GA, Garcia-Ojalvo J, Elowitz MB. 2010. *Cis*-interactions between Notch and Delta generate mutually exclusive signalling states. *Nature* **465:** 86–90.

Thompson A. 1997. An evolved circuit, intrinsic in silicon, entwined with physics. In *Evolvable systems: from biology to hardware*, Vol. 1259 of *Lecture notes in computer science* (ed. Higuchi T, et al.), pp. 390–405. Springer, Berlin, Heidelberg.

Van der Maaten L, Hinton G. 2008. Visualizing data using t-SNE. *J Mach Learn Res* **9:** 2579–2605.

Variano EA, Lipson H. 2004. Networks, dynamics, and modularity. *Phys Rev Lett* **92:** 188701.

Visscher PM, Medland SE, Ferreira MAR, Morley KI, Zhu G, Cornes BK, Montgomery GW, Martin NG. 2006. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet* **2:** e41.

Weinberg R. 2010. Point: hypotheses first. *Nature* **464:** 678.

Weinberg RA. 2014. Coming full circle—from endless complexity to simplicity and back again. *Cell* **157:** 267–271.

Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, Chu AY, Estrada K, Luan J, Kutalik Z, et al. 2014. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* **46:** 1173–1186.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, et al. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42:** 565–569.

You C, Okano H, Hui S, Zhang Z, Kim M, Gunderson CW, Wang YP, Lenz P, Yan D, Hwa T. 2013. Coordination of bacterial proteome with metabolism by cyclic AMP signalling. *Nature* **500:** 301–306.

Youk H, van Oudenaarden A. 2009. Growth landscape formed by perception and import of glucose in yeast. *Nature* **462:** 875–879.