EMBO reports

opinion

Taming the complexity of large models

Matthew Oberhardt & Eytan Ruppin

At its most basic, science is about models. Natural phenomena that were perplexing to ancient humans have been systematically illuminated as scientific models have revealed the mathematical order underlying the natural world. But what happens when the models themselves become complex enough that they too must be interpreted to be understood?

In 2012, Jonathan Karr, Markus Covert and colleagues at the University of California, San Diego (USA) produced a bold new biological model that attempts to simulate an entire cell: iMg [1]. iMg merges 28 sub-modules of processes within *Mycobacterium genitalium*, one of the simplest organisms known to man. As a systems biology big-data model, iMg is unique in its scope and is an undeniable paragon of good craft. Because it is probable that this landmark paper will soon be followed by other whole cell models, we feel it is timely to examine this important endeavour, its challenges and potential pitfalls.

Building a model requires making many decisions, such as which processes to glaze over and which to reconstruct in detail, how many and what kinds of connections to forge between the model's constituents, and how to determine values for the model's parameters. The standard practice has been to tune a model's parameters and its structure to a best fit with the available data. But this approach breaks down when building a large whole cell model because the number of decisions inflates with the model's size, and the amount of data required for these decisions to be unequivocal becomes huge. This problem is fundamental, not merely technical, and is rooted in the principle of frugality that underlies all science: *Occam's razor*.

The problem posed by Occam's razor is that there are vastly more potential large models that can successfully predict and explain any given body of data than there are small ones. As we can tweak increasingly complex models in an increasing number of ways, we can produce many large models that fit the data perfectly and yet do not reflect the cellular reality. Even if a model fits all the data well, the chance of it happening to be the 'correct' model—in other words the one that reflects correctly the underlying cellular architecture and relevant enzymatic parameters—is inversely related to its complexity. A sophisticated large model such as iMg, which has been fitted to many available datasets, will certainly recapture many behaviours of the real system. But it could also recapture many other potentially wrong ones.

How do we test a model's correctness in the sense just mentioned? The intuitive way is to make and test predictions about previously uncharted phenomena. But validating a large biological model is an inherently different challenge than the common practice of "predict, test and validate" customary with smaller ones. Validation using phenotypic 'emerging' predictions would require such large amounts of data that it would be highly inefficient and costly at this scale, especially as many of these predictions will turn out to be false leads, with negative results yielding little insight. Rather, the correctness of a whole-cell model is perhaps best validated by using a complementary paradigm: direct testing of the basic decisions that went into the model's construction. For example, enzymatic rate constants that were fitted in order to make the model behave properly could be experimentally scrutinized for later versions. Performing extensive sensitivity analyses and incorporating known confidence levels of modelling decisions, or harnessing more advanced methods such as 'active learning' should all be used in conjunction to determine which parameters to focus on in the future. The process of validating a large model should thus be viewed as an ongoing mission that aims to produce more refined and accurate drafts by improving low-confidence areas or gaps in the model's construction. Step by step, this paradigm should increase a model's reliability and ability to make valid new predictions.

An open discussion of the potential pitfalls and benefits of building complex biological models could not be timelier, as both the EU and the US have just committed more than a combined 1.4 billion dollars to explicitly model the human brain. Massive data collection and big data analysis are the new norm in most fields, and big models are following closely behind. Their cost, usefulness and application remain open for discussion, but we certainly laud the spirit of the effort. For what is certain is this: only by building these models will we know what usefulness we can attribute to them. Paraphrasing Paul Cezzane, these efforts might be indeed justified and worthy, so long as one is "more or less master of his model".

ACKNOWLEDGEMENTS

We thank Allon Wagner for helpful comments. Matthew Oberhardt is funded by the Whitaker Foundation and the Dan David Fellowship, and Eytan Ruppin is funded by the EU FP7 Microme project, the Israeli Science Foundation (ISF), the Israeli Centres of Research Excellence (I-CORE) and the McDonnell Foundation.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCE

1. Karr JR *et al* (2012) *Cell* **150:** 389–401

Matthew Oberhardt and Eytan Ruppin are at the School of Computer Sciences and Sackler School of Medicine, Tel Aviv University, Israel.
Matthew Oberhardt is additionally in the Department of Molecular Microbiology and Biotechnology, Faculty of Life Sciences, Tel Aviv University, Israel.
E-mails: mattoby@gmail.com; ruppin@post.tau.ac.il

EMBO *reports* (2013) **14,** 848; published online 6 September 2013; doi:10.1038/embor.2013.145