

HIGH-PERFORMANCE COMPUTING

Closing the Scientific Loop: Bridging Correlation and Causality in the Petaflop Age

Gary An

Published 21 July 2010; Volume 2 Issue 41 41ps34

Advances in computing capability offer the biomedical research community the prospect of creating simulations at a previously unimaginable scale. A diagnostic analysis of the underpinnings of the translational dilemma suggests that the current high-throughput, data-rich research environment has led to an imbalance in the relationship between determinations of correlation and the evaluation of causality. Here I describe the use of high-performance computing technologies to facilitate high-throughput hypothesis evaluation combined with an evolutionary paradigm for the advancement of scientific knowledge. This combination provides a roadmap for closing the scientific loop between correlation and causality, a necessary step if translational endeavors are to succeed.

Petaflop:

Definition, 1 quadrillion (“1” followed by 15 zeros) floating point operations per second. *Usage*, Metric for supercomputer performance.

The petaflop barrier was broken on 26 May 2008 by IBM’s Roadrunner machine at Los Alamos National Laboratory with a speed of 1.026 petaflops. As of 16 December 2009, the world’s fastest supercomputer is the Cray XT5 Jaguar machine at Oak Ridge National Laboratory, which has a measured speed of 1.75 petaflops and a theoretical top speed of 2.3 petaflops (1).

Translation: A whole lot of computing power . . .

Biomedical research today faces a critical dilemma. As huge amounts of detailed information become available for increasingly finer levels of biological mechanisms, it has become more and more difficult to effectively translate basic mechanistic knowledge into clinically effective therapeutics (2). The daunting challenge of translating mechanistic knowledge across scales of biological organization is a critical step in the development and evaluation of interventions for complex diseases such as sepsis, cancer, obesity, and autoimmune conditions. Sophisticated analytical methods aid in unraveling complexity by identifying patterns of relationships between increasingly dense data sets. The growing reliance and emphasis on

these methods raises the following question: Do researchers believe that, with enough data, correlation can be conflated with causality, and thus high-volume data analysis represents the future of science? Or are we at the point in the evolution of science where there simply is no better option? I propose that the currently perceived condition results from imbalances in the iterative loop of the scientific process, arising from differential advances in some aspects of the process versus others. A diagnostic approach to the translational dilemma can aid in the recognition of where deficits lie and guide the direction of future efforts. The results of such a diagnosis are presented below as a series of assertions regarding biomedical science, followed by a proposed approach for addressing the translational challenge within the context of the scientific method—the cycle of observation, experimentation, and interpretation that results in a set of predictive beliefs that constitute scientific knowledge.

THE ASSERTIONS

Assertion 1: Biological knowledge will always be incomplete. The goal of biology is not completeness of description or the striving for an ontological truth, but rather sufficiency of explanation. Sufficiency implies a contextual goal: How much do I need to know to accomplish what I want to do? The modeling and simulation community calls this “defining the experimental frame”: a specification of the conditions under which the system is observed and experimented with, thus determining constraints on what

can be interpreted from a result (3). Sufficiency also implies “trust” on the part of the researcher, given a particular experimental frame: How do I establish trust in the “truth” (albeit limited and constrained) of my knowledge? This point deserves emphasis, because it means that dealing with the incompleteness of knowledge is something we already do all the time! The need to deal with incompleteness through the determination of sufficiency is intrinsic to any research endeavor, but particularly so in translational research, which has as its eventual goal the improvement of public health through the implementation of acquired and interpreted knowledge that can be trusted.

Assertion 2: The ability to control a pathophysiological process requires an inference of mechanistic causality with respect to the biological process being targeted. This is a seemingly obvious but critical fact: If we hope to effectively intervene in and exercise control over a system, we must believe that the system operates on mechanisms that represent cause and effect. Imputed mechanisms of causality exist within, and are specific to, a defined experimental frame. Establishing trust in hypothesized causality is the goal of the scientific method. As a brief review and reference point, a schematic of the iterative cycle of science is seen in Fig. 1. The informed evaluation of any proposed intervention must incorporate some hypothesis of mechanistic causality in the system being targeted.

Assertion 3: The current translational dilemma results from an imbalance in the scientific cycle. Advances in technology have led to a significant increase in the flow through the scientific cycle, leading to potential bottlenecks: More data are being acquired, those data need to be mined and interpreted to suggest hypotheses, and those hypotheses need to be evaluated. Areas of ongoing research and development in bioinformatics can be linked to attempts to address these bottlenecks: data mining, curation, and sharing for high-throughput observations; data integration, pattern determination, and automated inference for hypothesis generation; and modeling and simulation for hypothesis testing (4, 5).

However, advances in technology and methodology do not necessarily enhance each arm of the scientific cycle equally and concurrently. Specifically, I propose that the development of methods for generating and capturing data (the observation step) and identifying patterns in those data (the

Section of General Surgery, Department of Surgery, University of Chicago, Chicago, IL 60637, USA.
E-mail: docgca@gmail.com

hypothesis-generation step) have outstripped the capacity to evaluate hypotheses derived from such correlative analyses. Under the traditional scientific method, mechanistic hypotheses derived from relationship patterns extracted from high-throughput data analysis would all be evaluated by experiments to generate new data to determine whether the suggested correlative patterns could be trusted as potential causal mechanisms. Unfortunately, the complexity and dimensionality (in terms of multiple factors and variables) of current biomedical problems preclude this approach. Therefore, an imbalance exists in our ability to address the various process bottlenecks in the current high-throughput environment, in which emphasis on the initial acquisition, management, and interpretation of data has superseded attention to evaluating the transition from correlation to causality (Fig. 2).

Although such an imbalance may be a necessary and natural component of the overall evolution of science (that is, it is reasonable to suppose that developments in data acquisition and management would precede augmentation of hypothesis testing), there is a potential danger that scientists begin to rely on the most sophisticated tools available at the expense of reestablishing balance in the cycle, leading to gridlock and stagnation [the current translational dilemma? (2)]. Therefore, the current state of affairs can be described as such: If a sophisticated data-mining, pattern-identification algorithm was used to identify the correlation, then attempting to reconstruct a traditional experiment to represent the system and test the hypothesis is an intractable challenge. However, data-driven computational analysis cannot be used to evaluate mechanistic causality in a hypothesis. Although correlative patterns may provide the foundational basis of causal hypothesis development, the scientific method mandates an additional step: experimental evaluation of causality. If the hypothesis evaluation/testing bottleneck evident in Fig. 2 results from the reliance on traditional stepwise experimental procedures (one experiment, one variable, one evaluation of causality), then it stands to

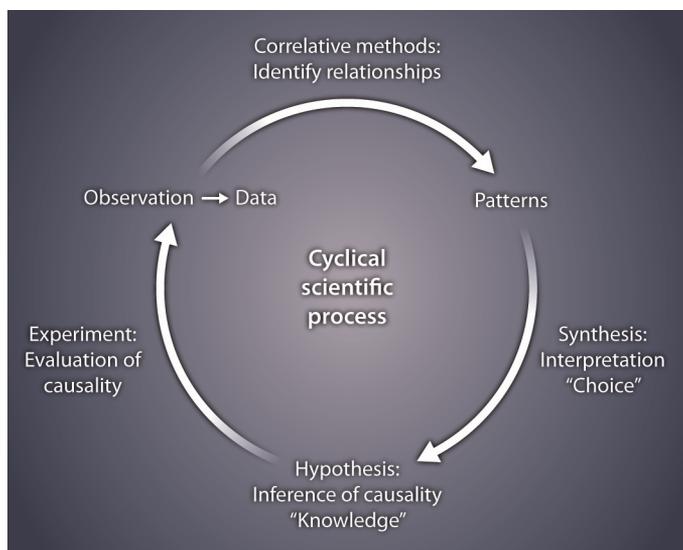


Fig. 1. Discovery channel. Within the iterative loop of the scientific process, data are produced by either observation or experiment. These data are then analyzed to detect putative relationships, which are used to form mechanistic hypotheses. This last step converts data into knowledge through interpretation. However, because correlation does not equal causality, the hypothesis must be subjected to an experiment, which is then evaluated statistically to establish trust in the causality represented within the hypothesis.

reason that the goal of scientific method development would be an attempt to increase the throughput potential of hypothesis evaluation. Computational advances, both in modeling and simulation methods and in hardware capability, represent the solution.

Assertion 4: High-throughput multiscale hypothesis evaluation requires computational representation of mechanistic hypotheses (instantiating thought experiments). All researchers construct mental models of the intellectual basis of their research, often represented in flow diagrams of the type so ubiquitous in biomedical papers. However, these diagrams are static and thus do not allow one to observe the consequences of relationships among components and mechanisms. The process of bringing these static diagrams to life has been termed executable biology (6), dynamic knowledge representation (7), or synthetic modeling and simulation (8). The current proposal emphasizes the development of high-throughput hypothesis testing/evaluation: the facilitation of the development of many models/simulations that represent a diversity of candidate hypotheses via the “democratization” of simulation construction. This will produce a distributed approach to discovering and testing what we know, building on those hypotheses suf-

ficient trust in understanding, and then moving toward intervention development. Thus, the target of sufficiency is raised again: At what point is a simulation sufficient for high-throughput hypothesis evaluation?

Fortunately, biological systems exhibit robustness and dynamic stability in a modular and multiscale fashion, where the macro-level behaviors are fairly resistant to micro-level fluctuations of underlying mechanisms: Even patients with chronic diseases function relatively normally with respect to the majority of their metabolic and homeostatic mechanisms, and the diseases themselves represent relatively stable dynamic systems. Therefore, assessments of macro-level behavior can be qualitatively evaluated with the goal of establishing face validity of a trans-scale mental model. Face validity asks that the simulation behave reasonably and be sufficiently accurate (9). Simulation

has established a sequence of tiers of validation that can be related to the needs of translational research (3, 9–12). These tiers represent steps toward increasing the resolution and predictive capability of simulations. Face validity is the first step. For biomedical research, this sequence corresponds to the movement from the discovery phase of science to the engineering phase of intervention development and testing. The key point is that you cannot skip steps; requiring a simulation or hypothesis to pass through the various stages of testing reduces the likelihood of being on the wrong track. Given the intrinsic incompleteness of biological knowledge, the research community’s needs in terms of developing viable translational strategies should be primarily focused on the earliest stages of simulation and hypothesis evaluation: establishing face validity. The challenge is that it is virtually impossible to know, a priori, which micro-level mechanisms will be significant at the macro level, even given the standard of face validity.

Emphasis on the goal of sufficiency in the translational endeavor provides some guiding principles: (i) The level of detail in a simulation or model is guided by the planned intervention; the level of simulation resolution must be at least as detailed

as the proposed mechanism of the planned intervention. (ii) Validation checks should be sought for each simulation scale of resolution. These are directed at falsification, to eliminate hypotheses that are clearly wrong and continue with hypotheses that are plausible; the goal is sufficiency, not “truth.” And (iii) information from productive falsification—the identification and explicit description of the reason for failure—is used to improve the simulation (heuristics). The high-throughput approach suggests that this process should be taking place in parallel, with multiple researchers, each with multiple hypotheses, exploring potential solutions. This is “massive induction” implemented at a community-wide level, occurring at multiple levels and within multiple modules representing different but interoperable defined experimental frames, matching the current structure of both biological systems and the scientific research community (13). At first, this prospect appears to be overwhelming, in terms of both implementation and analysis. However, developments in computing technology are moving toward providing this capability.

Assertion 5: Developing the capacity for computationally assisted, high-throughput hypothesis testing requires advances in methods to use high-performance computing (HPC) systems. As noted in the preamble, computing capability has reached previously inconceivable levels of sophistication. These advances in computational capability have historically followed two paths that are now merging: increasing individual computer power and increasing computer connectivity. The most powerful supercomputers are aggregates of millions of relatively simple processors (14, 15), and distributed systems [such as grid (16) and cloud computing (17)] are increasing in use and availability. A high-throughput hypothesis evaluation strategy for translational research needs to develop methodologies to use both technological avenues. For the purposes of this discussion, let us accept that there is already recognition of the importance and benefits of modeling and simulation (5–8) and also that many of these methods are relatively mature and capable, at

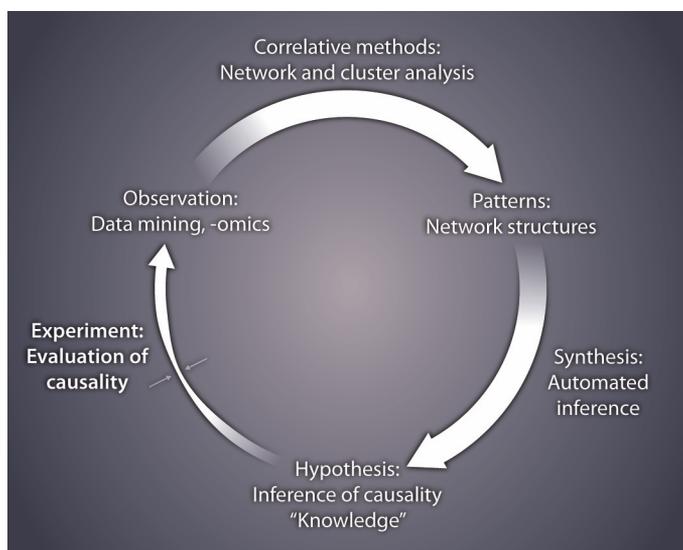


Fig. 2. Out of sync. This diagram depicts the current imbalance in the scientific process, in which high-throughput data collection has vastly increased the amount of existing data. These data are then processed using sophisticated correlation-establishing methods as a substitute for human insight. This leads to a large number of candidate hypotheses, and researchers need to exercise choice in selecting the components of their hypothesis. However, the causality inferred in these hypotheses cannot be tested sufficiently using traditional methods. This situation results in a bottleneck (two small arrows) in the scientific process at the point of causality evaluation.

least within defined-use cases. Let us also accept that there will be ongoing development in the refinement and expansion of methods. What trajectories of method development, then, are necessary to achieve the goal of high-throughput hypothesis evaluation? I propose three such primary areas: (i) facilitating the transfer of existing knowledge into simulation format, (ii) transfer of simulation types onto HPC platforms, and (iii) augmenting the ability to evaluate and interpret large-scale simulations and their output.

In terms of knowledge transfer, the experience of developing and managing bio-ontologies by the National Center for Biomedical Ontology (NCBO) (18) is particularly illuminating. Bio-ontology development follows a distributed paradigm, wherein multiple researchers construct specific ontologies, often tied to a particular domain and/or use. The NCBO provides a curated repository for these ontologies; use by the community drives the further refinement of the most-used ontologies. As a result, community-wide expertise is marshaled toward the advancement of methods for structured knowledge. Structured knowledge, in turn, provides an avenue into the modeling and simulation community. The conversion from structured knowledge

in current bio-ontologies to executable knowledge has already begun using a variety of simulation and modeling methods (19–21).

Substantial advancement in the capability to scale up the detail and resolution of the simulations must occur, and the use of HPC will play a critical role (3, 9–12, 22). With petaflop computing capability, simulations of previously inconceivable computational overhead now become feasible. This is vital, because the translation of knowledge regarding molecular interventions to the clinical context will require multiscale integration of hosts of simulation modules of increasing detail. Translational intervention engineering will require moving beyond abstract low-resolution models to high-fidelity simulations consistent with the processes described in the modeling and simulation arena (3, 10, 11).

It is inevitable that as simulations become more and more complex, analyzing their behavior becomes a computationally intensive task in and of itself; therefore, development must also occur in this area. However, there is an important difference between managing the complexity of a simulation versus the complexity of real-world biology: The simulation is constructed to be as transparent as is feasible. There are tools available for the evaluation of formally specified computer programs. Formal model checking (23), functional testing (24), and machine learning methods in the analysis of simulated hypotheses greatly increase our ability to determine where the gaps in knowledge and understanding may lie.

AN EVOLUTIONARY PARADIGM FOR TRANSLATIONAL SCIENCE IN THE PETAFLAP AGE

Currently, “data” is king, but data alone are not knowledge, and science is ultimately the advancement of knowledge. Highly complex, data-dense, multiscale problems call for a similarly high-dimensional solution; it should be noted that the population-level dimension of the research community is an integral aspect to this approach. Consider the example depicted in Fig. 3. Min-

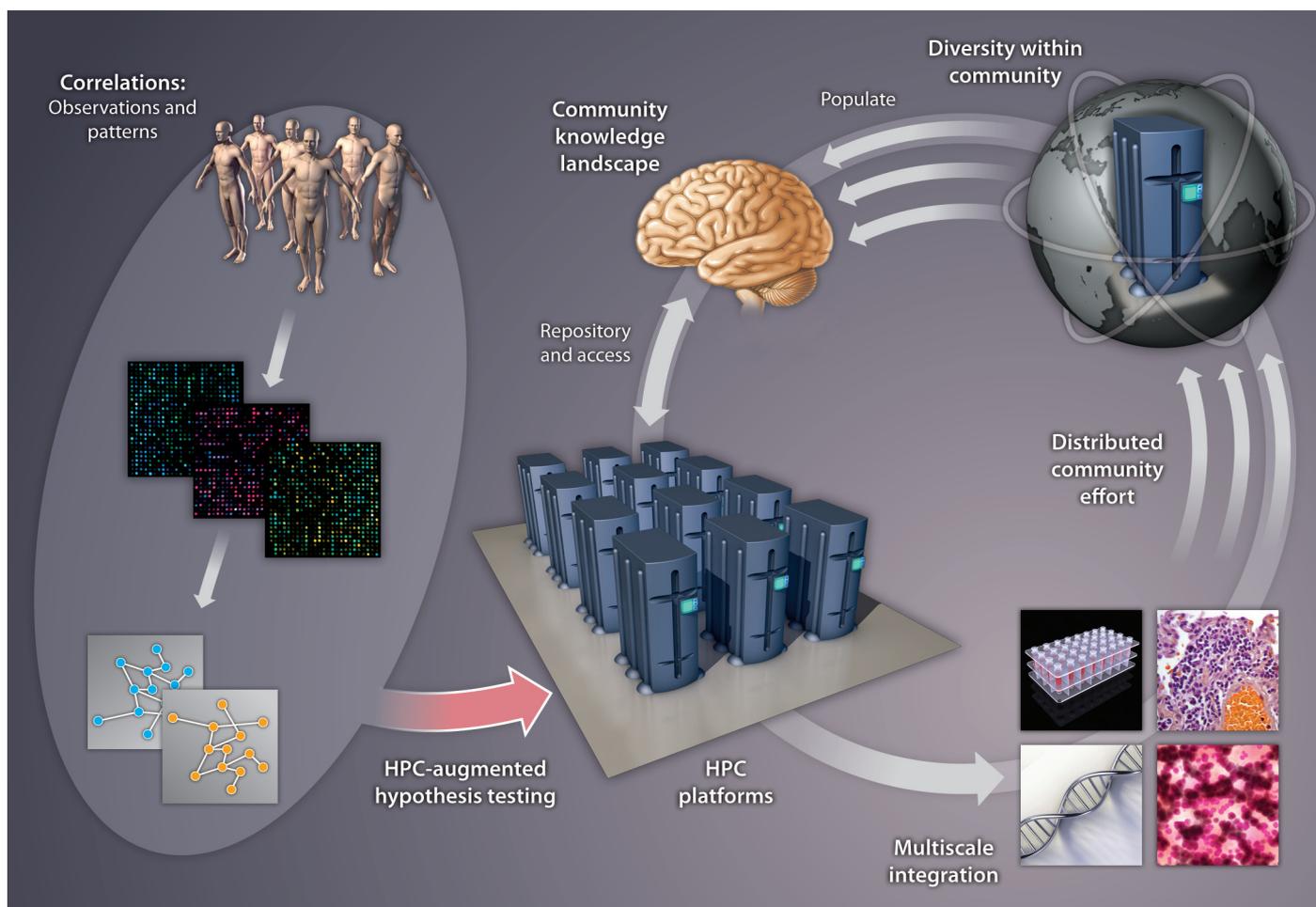


Fig. 3. Good causes. HPC can be used to augment high-throughput causality representation and testing. (From left to right, following arrows) Data mining and correlation identification lead to candidate relationship structures (patterns). HPC allows for parallel testing of multiple candidate causal hypotheses; plausible models are kept, nonplausible ones discarded. Concurrently, HPC allows for the instantiation of multiscale models (for example, of human disease physiology) constructed from modular subsystem models (for example, clinical chemistry parameters, gene and genome sequencing, and various kinds of imaging data). These research endeavors exist and function within a knowledge landscape that houses input from diverse scientific communities; HPC capabilities facilitate the underlying integrated repository framework: the “community knowledge landscape.”

ing of genomic data in a population of sepsis patients identifies n genes that are altered during a 28-day hospital course; network analysis demonstrates a network structure consisting of x primary, more highly connected nodes (“scale-free” topology); then metabolomic, mRNA, and cytokine samples identify approximately y to z compounds associated with each primary-node gene, providing candidate regulatory, signaling, and metabolic pathway structures. However, the set of possible hypothetical causal configurations is immense. So what are plausible causal pathway structures? High-throughput hypothesis evaluation implemented on HPCs allows the instantiation, in parallel, of multiple hypotheses and identifies the causal structures that can (i) recapitulate the original data and then (ii) be used as virtual

experimental platforms to evaluate new experiments. Furthermore, different pathways predominate in different cell types (inflammatory, endothelial, and epithelial cells, etc.), and these cell types are organized into different tissues and organs, all of which have specific behaviors in particular states of disease (shock, multiple-organ failure and recovery, etc.). As multiscale hypotheses arise to describe the various contributions and configurations of these factors and components, the petaflop capability of HPC platforms offers the possibility that these multiscale hypotheses can be instantiated with sufficient execution speed and diversity, so that a population of virtual experiments can be practically generated and performed, ranging from molecular biology to clinical trials.

This process is not confined to a single researcher or lab; rather, there is a community-level repository of this knowledge and these models, where different labs have simulation modules of either competing or complementary hypotheses and areas of focus. Collaborations consist of using and building on existing modules in the community repository. Some of these modules will remain relatively abstract and be cycled back for the task of discovery; others will be refined toward engineering-grade simulations for intervention development. Modules and models that are deemed useful by sufficient numbers of the community will persist; those that are not will fall by the wayside. High-throughput flow of new data feeds the scientific cycle, leading to continued and iterative modification of

the community-level knowledge landscape. All the necessary components of evolution are present: diversity, varying fitness, selection, and reproductive success (manifested as persistence). Diversity in science can be seen in community knowledge represented by the multiplicity of researchers' hypotheses. The fitness of a hypothesis is based on its ability to stand the scrutiny of experimental validation. Surviving hypotheses are those that are sufficient to explain their current data environment; if the data environment shifts, then the determination of fitness changes, and new hypotheses arise to populate the knowledge landscape. Tying the process of science to a theory—evolution—that is as fundamental and robust as exists in science would seem to be a favorable strategy.

Computational augmentation of causality evaluation increases the dimensionality of the testing and selection capacity and represents a high-throughput means of breaking the bottleneck in the current data- and correlation-dependent scientific environment. If the scientific method is to be preserved in today's complex, data-dense, multiscale environment, the strategy presented in this article must eventually occur. The advent of the petaflop age offers the promise that such a strategy is possible if given priority.

REFERENCES AND NOTES

1. Top500 Supercomputer Sites, <http://www.top500.org/> (accessed 16 December 2009).
2. *Innovation or Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products* (U.S. Food and Drug Administration, Washington, DC, 2004), <http://www.fda.gov/ScienceResearch/SpecialTopics/CriticalPathInitiative/CriticalPathOpportunitiesReports/ucm077262.htm>.
3. B. P. Ziegler, H. Praehofer, T. G. Kim, *Theory of Modeling and Simulation* (Elsevier, San Diego, ed. 2, 2000).
4. *Catalyzing Inquiry at the Interface of Computing and Biology*, J. C. Wooley, H. S. Lin, Eds. (National Academies Press, Washington, DC, 2005), pp. 35–56.
5. *Catalyzing Inquiry at the Interface of Computing and Biology*, J. C. Wooley, H. S. Lin, Eds. (National Academies Press, Washington, DC, 2005), pp. 117–204.
6. J. Fisher, T. A. Henzinger, Executable cell biology. *Nat. Biotechnol.* **25**, 1239–1249 (2007).
7. G. An, Introduction of an agent-based multi-scale modular architecture for dynamic knowledge representation of acute inflammation. *Theor. Biol. Med. Model.* **5**, 11 (2008).
8. C. A. Hunt, G. E. Ropella, T. N. Lam, J. Tang, S. H. Kim, J. A. Engelberg, S. Sheikh-Bahaei, At the biological modeling and simulation frontier. *Pharm. Res.* **26**, 2369–2400 (2009).
9. R. C. Kennedy, X. Xiang, T. F. Cosimano, L. A. Arthurs, P. A. Maurice, G. R. Madey, S. E. Cabaniss, Verification and validation assessment of simulation models, in *Annual Conference of the North American Computational Social and Organizational Sciences* (NAACSOS, Notre Dame, IN, 2006).
10. O. Balci, A methodology for certification of modeling and simulation applications. *ACM Trans. Model. Comput. Simul.* **11**, 352–377 (2001).
11. O. Balci, in *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice*, J. Banks, Ed. (Wiley, New York, 1998), pp. 335–396.
12. R. G. Sargent, Verification and validation of simulation models, in *Winter Simulation Conference 1998* (Association for Computing Machinery, New York, 1998), vol. 1, pp. 121–130.
13. G. An, Concepts for developing a collaborative in silico model of the acute inflammatory response using agent-based modeling. *J. Crit. Care* **21**, 105–110; discussion 110–111 (2006).
14. A. Gara, M. A. Blumrich, D. Chen, G. L.-T. Chiu, P. Coteus, M. E. Giampapa, R. A. Haring, P. Heidelberger, D. Hoenicke, G. V. Kocpsay, T. A. Liebsch, M. Ohmacht, B. D. Steinmacher-Burow, T. Takken, P. Vranas, Overview of the Blue Gene/L system architecture. *IBM J. Res. Dev.* **49**, 195–212 (2005).
15. Cray, ORNL's Jaguar XT5 Supercomputer, <http://www.cray.com/Products/XT/ORNLJaguar.aspx> (accessed 16 December 2009).
16. Folding@home, <http://folding.stanford.edu/English/Main> (accessed 16 December 2009).
17. I. Foster, Y. Zhao, I. Raicu, S. Lu, Cloud computing and grid computing 360-degree compared, in *Grid Computing Environments Workshop 2008, Austin, TX* (IEEE, Los Alamitos, CA, 2008).
18. National Center for Biomedical Ontology, <http://www.bioontology.org/>.
19. T. Takai-Igarashi, Ontology based standardization of Petri net modeling for signaling pathways. *In Silico Biol.* **5**, 529–536 (2005).
20. G. An, Dynamic knowledge representation using agent-based modeling: Ontology instantiation and verification of conceptual models. *Methods Mol. Biol.* **500**, 445–468 (2009).
21. D. Shegogue, W. J. Zheng, Integration of the gene ontology into an object-oriented architecture. *BMC Bioinformatics* **6**, 113 (2005).
22. C. Deissenberg, S. van der Hoog, H. Dawid, EURACE: A massively parallel agent-based model of the European economy. *Appl. Math. Comput.* **204**, 541–552 (2008).
23. E. M. Clarke, E. A. Emerson, A. P. Sistla, Automatic verification of finite-state concurrent systems using temporal logic specifications. *ACM Trans. Prog. Lang. Sys.* **8**, 244–263 (1986).
24. F. Ipaté, M. Holcombe, A method for refining and testing generalised machine specifications. *Int. J. Comput. Math.* **68**, 197–219 (1998).
25. This work was supported in part through NSF grant 0830-370-V601. I thank C. Anthony Hunt and G. Ropella for discussions that inspired this article. **Competing interests:** G.A. is a consultant for Immunetrics, Inc.

10.1126/scitranslmed.3000390

Citation: G. An, Closing the scientific loop: Bridging correlation and causality in the petaflop age. *Sci. Transl. Med.* **2**, 41ps34 (2010).