

Verification of systems biology research in the age of collaborative competition

Pablo Meyer¹, Leonidas G Alexopoulos², Thomas Bonk³, Andrea Califano⁴, Carolyn R Cho⁵, Alberto de la Fuente⁶, David de Graaf⁷, Alexander J Hartemink⁸, Julia Hoeng³, Nikolai V Ivanov³, Heinz Koepl⁹, Rune Linding¹⁰, Daniel Marbach¹¹, Raquel Norel¹, Manuel C Peitsch³, J Jeremy Rice¹, Ajay Royyuru¹, Frank Schacherer¹², Joerg Sprengel¹³, Katrin Stolle³, Dennis Vitkup⁴ & Gustavo Stolovitzky¹

Collaborative competitions in which communities of researchers compete to solve challenges may facilitate more rigorous scrutiny of scientific results.

Systems biology aims to provide a mechanistic understanding of biological systems from high-throughput data. Besides its intrinsic scientific value, this understanding will accelerate product design and development, facilitate health policy decisions and may reduce the need for long-term clinical trials. For this to happen, the knowledge generated by systems biology has to become suffi-

ciently trustworthy for the empirical approach underlying long-term clinical trials to be supplanted by an approach in which mechanism and mechanistic understanding is a driver for decisions. This raises fundamental questions of how to evaluate the veracity of predictions from systems biology models and how to construct mechanistic models that best reflect biological phenomena—questions that are of interest to both academia and industry.

High-throughput verification of systems biology

In 2009, a report¹ from the US National Academy of Sciences (Washington, DC) highlighted four areas where biology could make major contributions: food production, improvement of human health, optimized biofuels and ecosystem restoration. Addressing these challenges requires not only multidisciplinary teams to analyze high-throughput quantitative data, but also verification of the conclusions from such analyses.

One of the obvious steps in raising the confidence of high-throughput data sets is to have better experimental and analytical techniques that yield accurate and reproducible data with known error rates. For example, verification of mass spectrometry proteomic measurements has proven difficult because the measurements can depend strongly on sample preparation, the method of detection and the biological context in which the measurements were made. One approach to address this issue has been the creation of databases such as Peptide Atlas², a genome-mapped library of peptides

derived from liquid chromatography tandem mass spectrometry proteomics experiments in multiple organisms that lends itself to easy navigation using software tools.

Another example of recent efforts to ensure data quality and reproducibility is the area of genome-wide association studies (GWAS), where researchers take an unbiased survey of common single-nucleotide polymorphisms (SNPs) across the genome and look for alleles whose presence correlates with phenotypes such as disease. Hundreds of gene candidates have been found in just a few years, although most have only a modest effect³. The difficulty is that slight differences in the genetic backgrounds of different populations or unknown pairs of relatives in a sample introduce tiny statistical shifts that pose the risk of appearing significant for some of the millions of SNPs analyzed. In response to these difficulties, researchers have adopted a well-defined, quality-control process that can be applied to new data using readily available software tools. Also, many journals are starting to require replication of results for publication of GWAS papers, and in the best scenario, another research group replicates the association study in a different cohort with a similar phenotype⁴.

The complex networks that translate genotype into phenotype are also highly sensitive to biological context and environmental influences. Typically, context and environment are mediated by signaling networks, for example, through the action of protein kinases. To verify predictions, it is necessary to understand how

¹IBM Computational Biology Center, Yorktown Heights, New York, USA. ²National Technical University of Athens, Athens, Greece. ³Philip Morris International R&D, Neuchâtel, Switzerland. ⁴Center for Computational Biology and Bioinformatics, Department of Biomedical Informatics, Columbia University, New York, USA. ⁵Modeling and Simulation, Merck & Co., Rahway, New Jersey, USA. ⁶Center for Advanced Studies, Research and Development in Sardinia (CRS4), Laboratorio di Bioinformatica, Parco tecnologico della Sardegna, Pula, Italy. ⁷Selventa, Cambridge, Massachusetts, USA. ⁸Duke University, Durham, North Carolina, USA. ⁹ETH Zurich, Zurich, Switzerland. ¹⁰Cellular Signal Integration Group (C-SIG), Center for Biological Sequence Analysis (CBS), Department of Systems Biology, Technical University of Denmark (DTU), Copenhagen, Denmark. ¹¹Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology (MIT) and Broad Institute of MIT and Harvard, Cambridge, USA. ¹²Biobase GmbH, Wolfenbuettel, Germany. ¹³IBM Life Sciences Division, Zurich, Switzerland. e-mail: gustavo@us.ibm.com

a network functions and to analyze its dynamical changes under certain conditions. A reliable source of quantitative data allows such predictions through the probabilistic integration of different sources of evidence, as in the case of NetPhorest⁵ (which creates an index to measure the specificity of protein kinases) or NetworKIN⁶ (which predicts the interactions between kinases and the substrate proteins they phosphorylate using cellular contextual information). When assessing the performance of such biological classifiers or predictions from models, it is essential to design experiments that reproduce the biological context as closely as possible, and to make use of independent data to corroborate the predictions. This is also the case for the underlying proteomics data.

Although computational methods and high-throughput experiments can be used to map interactions at the genome-wide level, they are often characterized by substantial error rates. Thus, many of the predicted interactions may be incorrect. Critically, these errors and their sources can be identified, quantified and corrected as our knowledge of the underlying system grows. Notably, for many applications it is not crucial that all predicted interactions be correct. For example, for the purpose of identifying master regulators—genes that orchestrate regulatory programs in transcriptional regulatory networks—it does not matter whether the researcher knows which transcription factor–target interactions are correct because, if a sufficiently high percentage of the interactions are correct, then in all likelihood the correct regulator will be predicted⁷.

Traditional approaches to validation are not particularly amenable to testing hundreds or thousands of potential interactions. However, verifying all the detailed mechanisms conjectured to underlie a biological system may be unnecessary until the model predicts something biologically important. Moreover, a hierarchical validation could be possible, where many predictions are validated at low resolution, and a few of them then investigated in greater detail.

Limitations of peer review for validation

Traditional peer review is widely considered to be one of the most important mechanisms for quality control of scientific papers. Nevertheless, as the number of published papers increases, the peer review system is under increasing strain. Indeed, it has been estimated from PubMed that in the past decade the growth rate of scientific publications was 5.6% per year, or equivalently, a doubling time of 13 years⁸. This results in increased burdens on peer reviewers who get little reward for their efforts. Furthermore, it is questionable whether

the peer review system can objectively assess the quality of the high-throughput data and the validity of the sophisticated analyses and interpretations that nowadays pervade systems biology.

Web-based publishing has created new mechanisms for gauging the reactions to a paper in the same journal in which it is published. The discussions and opposing opinions about an article considerably enrich it, or at least they could if they were more frequently used. In general, participation in open discussions of papers has not attracted the interest of many scientists, except for a few controversial papers. One way to improve reader feedback on journal websites may be to use a unique author identifier⁹ that is assigned to researchers early in their career so that their online comments and reviews can be taken into account during evaluations, in addition to their reviewed publications.

The proliferation of publications, which is a sign of the faster pace of discovery, may also dilute important discoveries as they may be split across several papers. One of the responses to this reality has been the creation of annotated biological databases (pioneered by SwissProt for over 20 years) based on the peer-reviewed literature. For example, Biobase (<http://www.biobase-international.com/>) annotates literature data, having processed some 150,000 references on the human proteome; curation is done by trained and paid curators. A similar effort involving massive human curation is being pursued by Ingenuity (<http://www.ingenuity.com/>). But, comparable with some of the subjectivity that exists in the peer review process, personal biases and different conclusions may be drawn for the same paper, even by highly skilled, rigorously selected curators who follow standard operating procedures.

As the explosive growth of biomedical data strains the capacity of human curation, computational methods to mine the literature are becoming increasingly important¹⁰. But automatic text mining has its own weaknesses, such as the difficulty of extracting information from figures or tables, and the ambiguities of interpretation inherent in natural language. Biological databases, whose information is usually subjected to some human curation, contain data and annotations that should be scrutinized for accuracy. In genomic databases, inadvertent annotation errors can be propagated when the putative function of a gene is inferred based on sequence homology. For example, current methods for biochemical annotation of metabolic pathways, especially for microbes, are primarily based on sequence homology and can be inaccurate because most annotations do not provide a quantification of

confidence of the homology. New methods need to be developed to police these errors in databases and avoid the propagation of incorrect information¹¹.

Naturally, it would be better to prevent mistakes from entering into the literature or the databases in the first place. But peer review and human curation can only address some of the inaccuracies that are contained in papers or databases. Peer reviewers mainly judge the suitability of data collection protocols, accuracy of inferences and ideas, innovation, the logic of the argument and the consistency of the material being reviewed. However, it is often unfeasible or difficult for reviewers to assess the quality of data itself or the performance of the analytical methodologies described in a manuscript. This is in part due to the lack of a rigorous characterization of error rates and data quality in manuscripts. Thus, a methodology is needed for verifying the results and claims in systems biology.

The power of crowds

In this respect, crowdsourcing—engaging an interested community to collaboratively solve a problem—may be a fruitful strategy for assessing the quality of analyses and predictions from high-throughput data. As one example of such an approach, blogs and tweets can gather sizable amounts of comments on controversial papers¹². For instance, as the popular media were covering a paper that identified genomic loci predicting human lifespan with 77% accuracy, scientific bloggers were already raising doubts about the methodology of the paper and the lack of rigor of its results, which blunted enthusiasm for the publication and led to its eventual retraction¹³.

Crowdsourcing has also been used to assess the validity of research in academic efforts such as CASP (The Critical Assessment of Protein Structure Prediction¹⁴; <http://predictioncenter.org/>), CAPRI (The Critical Assessment of Prediction of Interactions; <http://www.ebi.ac.uk/msd-srv/capri>), BioCREATIVE (<http://www.biocreative.org/>) and DREAM (The Dialogue on Reverse Engineering Assessments and Methods¹⁵; <http://www.thedream-project.org/>) as well as commercially organized assessments like Kaggle (<http://www.kaggle.com/>) and Innocentive (<http://www.innocentive.com/>). These undertakings are organized around ‘challenges’ in which an interested community competes to verify methodologies against carefully chosen benchmarks. We briefly discuss below two projects that illustrate the power of this approach, CASP (a pioneering project on collaborative competition) and DREAM (which deals with the assessment of methods for systems biology).



CASP is a contest begun in 1994 to rank the performance of methods for predicting the three-dimensional structure of proteins based on their amino acid sequence. It is the first of the many biological community-based assessment efforts to emerge. The nine CASP competitions to date have uncovered significant stumbling blocks in the field of protein structure prediction, and they have enabled notable progress.

DREAM is a project designed to assess model predictions and pathway inference algorithms in systems biology. Like CASP, DREAM is structured in the form of challenges presented to the community, comprising open problems whose solutions (the 'gold standards') are known to the organizers but not to the participants. Participants submit their predictions of the solutions, which are evaluated by the organizers and eventually discussed in a conference. After the conference, all the data, predictions and gold standards are openly available to the community. This experience has shown that a rigorous scrutiny of scientific research based on community involvement is possible. The outcomes of the DREAM challenges highlight areas in which clear advances in systems biology have been made or need to be made.

Several of the challenges posed by DREAM address the problem of 'network inference'. In these challenges, teams of researchers try to infer gene-regulatory or signaling networks from gene expression or phosphoproteomic profiles undergoing various perturbations.

This is a difficult problem because currently no true gold standard exists for real biological networks. Data simulated with mathematical models that are designed to be as biologically plausible as possible can be used, as simulated data assure a systematic, rigorous assessment¹⁶. But the use of simulated data does not ensure that the challenge is necessarily realistic¹⁷. Many different methods, including regression, mutual information, correlation, Bayesian networks and others¹⁸, have been used to address this challenge. Importantly, combining individual predictions results in a solution that is highly robust and usually the most accurate, demonstrating the need for tackling complex problems as a community^{18,19}.

For both CASP and DREAM, as well as for most similar efforts, the goal is not about finding a single best method, but rather, reaching a better understanding of the strengths and weaknesses of these methods to enable progress in their respective disciplines.

Meeting the needs of industry

In view of the limited ability of peer review to assure the validity of complex scientific results in the area of systems biology (Fig. 1), and recognizing the power of communities to assess methodological aspects of scientific research, researchers at IBM and Philip Morris International (PMI; Neuchâtel, Switzerland) have been collaborating on a vision for quality assurance in systems biology research. Although industry shares many of the same

needs for validation as academia, a methodology for verifying research is needed in the industrial setting that recognizes both speed and protection of proprietary data constraints, as well as the importance of market considerations and consumer protection. IBM and PMI have proposed a scheme called IMPROVER (industrial methodology for process verification of research; Box 1).

Applying this methodology first requires identifying the building blocks of a research workflow. Building blocks are basically small pieces of a big research program. Some might involve generating biological measurements, others analyzing data. The validity of these measurements can be assured with quality-assessment and best-practice processes that are familiar to industry. The idea behind IMPROVER is to test each key method at crucial junctures of a research workflow by posing challenges designed to see whether or not the process works at the necessary level of accuracy (Box 1 and Fig. 2). The challenges can be internal to a company, or if they are of interest to a broader community that may benefit from its participation, they can be public, similar to DREAM or CASP. For such external challenges, the organizers will need to establish the same sorts of conditions that have made existing programs successful. In particular, for independent researchers to participate, they will need incentives, which could be recognition or co-authorship, monetary incentives (for instance, as used in the Netflix Prize²⁰) or access to high-value data or experimental validation efforts. The workflow model underlying this methodology reflects an engineering mindset more common in industry, where research is aimed at a concrete fixed goal, but not particularly adapted for completely open-ended discovery, as in academia.

IMPROVER could start a trend by which eventually even the academic community would ask for independent verification of its core technologies and methods. Today, independent verification relies mainly on government agencies whose criteria for assessment are not always transparent to the general public.

Finding a robust signature for disease diagnosis is an example of a challenge that might be of interest to the wider biomedical research community, as well as being essential in many industrial research workflows for stratification of populations, early detection of disease and personalized medicine. Pioneering work²¹ suggested that molecular classification of tumors, and by extension other diseases, could be more accurate than morphological classification. Similarly, success in predicting survival, disease progression and response to drugs could aid in stratifying patients and choosing

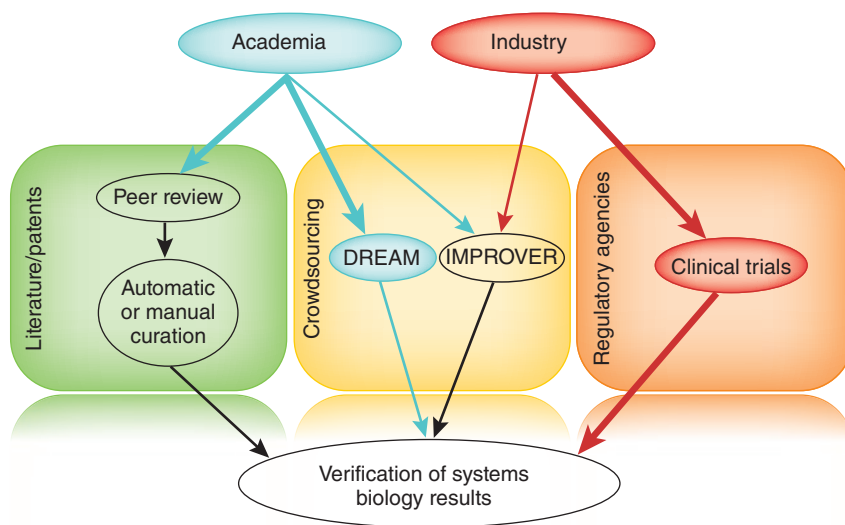


Figure 1 Current approaches to systems biology verification. Different paths for reaching systems biology verification are represented, both for academia (blue) and for industry (red). Black represents pathways common to industry and academia. The color of the rectangles represent the grounds on which the assessment of systems biology results are based: mostly on innovation (green), mostly robustness (orange) and both innovation and robustness (yellow). The thickness of the arrows represents the current predominant pathway.

Box 1 IMPROVER as a means of assessing complex processes in industrial research

The IMPROVER methodology verifies research workflows (top row; **Fig. 2**) by decomposing them into discrete building blocks (middle row; **Fig. 2**), which might represent a method, process or algorithm. IMPROVER is useful in an industrial research organization to assess the risk of a workflow by identifying problematic building blocks that may be, for instance, inaccurate or nonrobust. The methodology uses crowdsourcing to verify a building block by posing challenges (lower row; **Fig. 2**) to improve or derive solutions to that step in the workflow. In an internal challenge, verification is accomplished by a challenge within a company or organization. Participants' submissions are compared to a desired reference output (known as a gold standard) by a trusted referee who is blinded to the expected results. The building block is verified if the two results match within a predetermined difference criterion. The purpose of the example internal challenge (shown in blue) is to verify transcriptome data by comparing measurements with a reference data set of known quality to ensure sufficiently low noise level. In an external challenge, data are disseminated to participants outside the organization. Submissions are collected and compared to the reference data set by a trusted referee who is again blinded to the expected results. The purpose of the example external challenge (shown in red) is to verify whether discovered disease signatures in the transcriptome data are sufficient to predict known disease phenotypes.

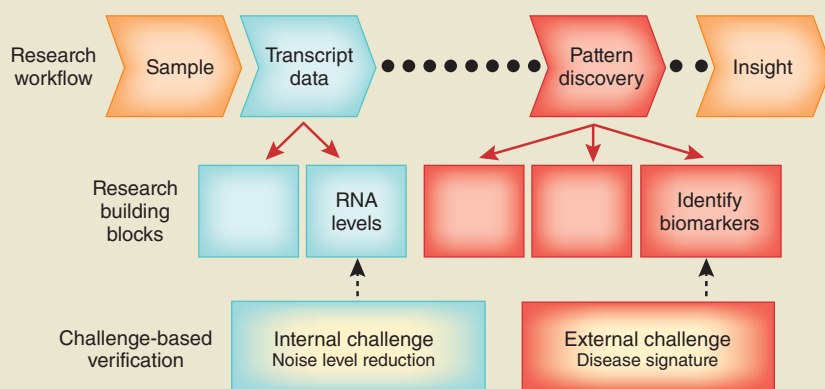


Figure 2 Example application of IMPROVER for verification of a plausible research workflow.

These challenges, as well as many others that can be envisioned, address the core interests of many industries. These industries could benefit from the power of crowds to find strategies to address their problems. In turn, the community will have the opportunity to try their methods on new data, participate in studies that address grand challenges in biomedical research and close the sometimes open loop between academia and industry.

Conclusions

The abundance of high-throughput and quantitative data in systems biology creates both opportunities and difficulties. In particular, although thousands of predictions may be generated, most are often left unverified. How worthwhile can these predictions be without methods for high-throughput verification? Several avenues to verification of systems biology results exist or are emerging in both academia and industry (**Fig. 1**).

In this article, we have proposed that systems biology results can be verified using community-wide challenges that test specific methodologies using the power of crowds. Assessing the results of these challenges needs to be done under a rigorous statistical framework. It is curious that little has been done in the area of verification of industrial research, especially as statistical quality control has enabled considerable improvements to industrial manufacturing. If challenge-based verification processes, such as IMPROVER, CASP and DREAM, become routine, it is likely that industrial research workflows will see increased efficiency in the generation of applied scientific results and decreased expense per verified result.

Challenge-based verification processes may also help cope with the explosive growth of scientific publications. This growth taxes the peer review system, especially in systems biology where assessments of the robustness of a complex methodology and the sanity of large data sets are often not performed during the peer review process. We argue that challenge-based verification of scientific results should ideally be done before submission to a reviewer. This could provide better scrutiny of results, because blind tests tend to eliminate some of the subjective bias of interpretation of results during peer review.

Finally, we should stress that an overcrowded field of scientific publications and a lack of systematic verification of systems biology predictions, although problematic, are consequences of something fundamentally positive, because they reflect the fact that science is moving at a fast pace. We hope that some of the specific solutions we have

treatments. For a signature to be robust, it will probably need to be more than just a single biomarker or gene signature. For example, it could be a set of master regulators of a tumor type⁷ or a combination of clinical data, gene expression data, pathway information and genomic structural variants^{22,23}. Integrative network markers and network structures and dynamics will increasingly become a primary focus for both detection and treatment of complex diseases. In this type of challenge, participants would be assessed on their ability to identify disease phenotypes based on gene expression data and, possibly, clinical information. The training set would perhaps not be given explicitly, with participants needing to rely on vast publicly available gene expression databases, such as the National Center for Biotechnology Information's Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo/>).

A second challenge that may find many

industrial applications would be the exploration of the limits of translation of data and conclusions from rodents to humans. The main scientific question here is how accurately observations from *in vivo* and *in vitro* rodent models can be translated to a human context. Participants would be given proteomics or expression profile changes in cultured cells, from a particular tissue from both rodents and humans, in response to an agent such as a drug. The challenge would be to predict the response in human cells to a new agent, on the basis of expression changes in rodent cells. One essential element in designing such a challenge is choosing the agent and the cell lines to give a diversity of perturbations that sufficiently cover the maximum number of biological processes. In addition to finding useful methodologies, the goal of this challenge would be to provide insight and understanding regarding the range of applicability of the translation concept.

outlined will help avoid false steps in our path toward a more predictive, quantitative and mechanistic understanding of biological systems.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at <http://www.nature.com/naturebiotechnology>.

ACKNOWLEDGMENTS

This paper was the result of vibrant discussions during the recent symposium “Critical Assessment of Systems Biology: Research Verification in the Age of Collaborative-Competition,” which took place in Zurich, Switzerland, on March 23 and 24, 2011. We thank S. Stadler, C. Haettenschwiler and C. Warmer for the organization of the symposium, D. Monroe for

a careful proofreading of the manuscript, R. Aebbersold, B. Schwikowski, S. Corthesy, K. Kozikis, L. Schilli and all the attendees who are not authors for their contributions as speakers or during the discussion sessions.

1. The US National Academy of Sciences. *A New Biology for the 21st Century* (National Academies Press, Washington, DC, 2009).
2. Zhang, Q. *et al. Genome Biol.* **9**, R93 (2008).
3. McCarthy, M.I. *et al. Nat. Rev. Genet.* **9**, 356–369 (2008).
4. Chanock, S.J. *et al. Nature* **447**, 655–660 (2007).
5. Miller, M.L. *et al. Sci. Signal.* **1**, ra2 (2008).
6. Linding, R. *et al. Cell* **129**, 1415–1426 (2007).
7. Carro, M.S. *et al. Nature* **463**, 318–325 (2010).
8. Larsen, P.O. & von Ins, M. *Scientometrics* **84**, 575–603 (2010).
9. Wolinsky, H. *EMBO Rep.* **9**, 1171–1174 (2008).
10. Harmston, N., Filsell, W. & Stumpf, M.P. *Hum. Genomics* **5**, 17–29 (2010).
11. Hsiao, T.L. *et al. Nat. Chem. Biol.* **6**, 34–40 (2010).
12. Mandavilli, A. *Nature* **469**, 286–287 (2011).
13. Sebastiani, P. *et al. Science* **333**, 404 (2010).
14. Moulton, J. *Curr. Opin. Struct. Biol.* **15**, 285–289 (2005).
15. Stolovitzky, G., Monroe, D. & Califano, A. *Ann. NY Acad. Sci.* **1115**, 1–22 (2007).
16. Marbach, D. *et al. J. Comput. Biol.* **16**, 229–239 (2009).
17. Cantone, I. *et al. Cell* **137**, 172–181 (2009).
18. Marbach, D. *et al. Proc. Natl. Acad. Sci. USA* **107**, 6286–6291 (2010).
19. Prill, R.J. *et al. PLoS ONE* **5**, e9202 (2010).
20. Tuzhilin, A. & Koren, Y. *ACM Digital Library* (ACM Press, New York, 2008).
21. Golub, T.R. *et al. Science* **286**, 531–537 (1999).
22. Erler, J.T. & Linding, R. *J. Pathol.* **220**, 290–296 (2010).
23. Tamayo, P. *et al. J. Clin. Oncol.* **29**, 1415–1423 (2011).