The inconvenience of data of convenience: computational research beyond post-mortem analyses

To the Editor: Over the last two decades researchers have witnessed an explosion in the amount and diversity of data collected in biological and medical studies. These data are often generated without the input of those who will later analyze it. Computational analyses are therefore, in the words of statistician Ronald Fisher, mostly performed 'post mortem'. We believe that a more efficient scientific process should use computational modeling based on previously acquired data to guide targeted data collection efforts.

We consider systematic data collection and model-driven data collection as distinct efforts. Large-scale systematic data collection efforts, such as TCGA, ENCODE, REMC, GTEx and the Connectivity Map, to name a few, have unquestionably led to important and actionable findings such as identifying treatment targets (https://cancergenome.nih.gov/researchhighlights/tcgainaction/tcga-data-used-for-loxo101-drug-development) and gaining insight into gene regulatory processes[1]. However, such data could have been even more useful. For example, in our own work on glioblastoma subtype discovery[2] , we could use only 46% of the TCGA samples because of missing measurements, reducing the power of the study. In another example, the fixed concentration levels of small-molecule compounds in the Connectivity Map were suboptimal for some compounds and cell contexts, leading to substantial batch effects[3].

DREAM Challenges, which harness the collective skills of computational biologists across the world to solve biological and medical problems using 'data of convenience', have illustrated the difficulties in this process[4–6]. For instance, in a DREAM challenge for predicting response to drugs in patients with rheumatoid arthritis, using the largest available collection of single-nucleotide polymorphism (SNP) data did not improve predictions over those obtained using the clinical predictors[5]. In a toxicogenetic challenge, genome-wide association study (GWAS) data by themselves were not predictive, but the results were markedly better when these were taken together with RNA-seq data, available for only 38% of the patients[4]. Finally, in a DREAM challenge assessing and improving drug sensitivity prediction algorithms, having data from many omics modalities did not provide an advantage over the use of gene expression data alone[6]. We concede that these situations may arise because some computational approaches are just not good enough for the task. However, the fact that none of several dozen independent expert teams were successful in solving the problems using the same data suggests that, instead, more or different kinds of data may be needed. The question then arises: How can one efficiently determine which data we *need to*, rather than *can,* measure to accelerate scientific discovery?

Hypothesis-driven experiments are common in the life sciences but tend to be small in scale. We argue that computational models, capable of generating targeted hypotheses that capture the complexity of biological systems, should be used to guide data collection. This offers the possibility not only of speeding up data collection but also of yielding better biological insights, thanks to the exploitation of more appropriate data. Recent successes in physics, such as the discovery of gravitational waves and the Higgs boson, illustrate the benefits of model-based experimentation very well. The biomedical field needs such examples of its own.

We firmly believe that computational biologists can contribute productively to model-driven experimental research. Models derived from more classical post-mortem data analysis should now guide the next wave of hypothesis generation, experimental design and data collection. To identify biomedical problems ready to be tackled, we have invited computational biologists from around the world to take part in the Idea DREAM Challenge (http://tinyurl.com/dreamidea). Participants were asked to propose biomedical research questions for which computational models have exploited available data to the limit and are ready to guide new data collection efforts to move the field forward. Through peer review and discussions among participants, we selected two winning ideas. We are now matching the winning participants with wet-lab researchers to generate the necessary data.

The first idea addresses the challenge of drug–target interaction mapping. The potential chemical space of drug-like compounds is thought to contain on the order of 10^{20} molecules, making exhaustive exploration infeasible. Furthermore, currently available bioactivity measurements vary greatly between labs and assay types, and hence are not yet sufficient to reliably guide the computational prediction of compound–target relationships at a large scale. One of the winning DREAM ideas proposed a model-guided experimental design and mapping effort to prioritize the most potent target selectivity experiments among the massive search space of compounds and their potential targets. Such targeted experiments, which will be predicted by computational models, are expected to offer a cost-effective alternative to the more systematic exploration efforts, effectively providing higher information content with the same amount of experiments.

The other winning DREAM idea tackles the problem of regulatory network inference, predicting which regulatory proteins control the expression of which target genes. The proposal is to systematically and iteratively collect multi-omic measurements under different genetic and environmental perturbations from both bulk populations and single cells. These data will be collected in a model-guided manner, in which the initial model is a consensus derived from published datasets to avoid duplication of experimental effort and enable maximal discovery. The resulting dataset will serve as a better gold standard to validate computational predictions from existing and new inference methods and will help identify the most informative datasets for regulatory network discovery.

We envision that the Idea DREAM Challenge is just the beginning of many more endeavors in which data analysts and computational biologists can be actively engaged in all stages of the scientific process. Model builders and experimentalists would benefit from working together to design better studies that will accelerate scientific discovery.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Chloé-Agathe Azencott[1–3], Tero Aittokallio[4,5], Sushmita Roy[6,7], DREAM Idea Challenge Consortium[8], Thea Norman[9], Stephen Friend[9], Gustavo Stolovitzky[10,11] & Anna Goldenberg[12,13]

[1]MINES ParisTech, PSL-Research University, CBIO–Centre for Computational Biology, Fontainebleau, France. [2]Institut Curie, Paris, France. [3]INSERM U900, Paris, France. [4]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Finland. [5]Department of Mathematics and Statistics, University of Turku, Finland. [6]Department of Biostatistics & Medical Informatics, University of Wisconsin–Madison, Madison, Wisconsin, USA. [7]Wisconsin Institute for Discovery, University of Wisconsin–Madison, Madison, Wisconsin, USA. [8]A list of members and affiliations follows. [9]Sage Bionetworks, Seattle, Washington, USA. [10]IBM Computational Biology Center, Yorktown Heights, New York, USA. [11]Department of Genetics and Genomics Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [12]Genetics and Genome Biology, SickKids Research Institute, Toronto, Ontario, Canada. [13]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada.

DREAM Idea Challenge Consortium:
Ankit Agrawal[14], Tero Aittokallio[4,5], Chloé-Agathe Azencott[1–3], Emmanuel Barillot[15], Nikolai Bessonov[16], Deborah Chasman[7], Urszula Czerwinska[15], Alireza Fotuhi Siahpirani[17], Stephen Friend[9], Anna Goldenberg[12,13], Jan Greenberg[18], Manuel Huber[19], Samuel Kaski[20,21], Christoph Kurz[19], Marsha Mailick[22], Michael Merzenich[23], Nadya Morozova[24,25], Arezoo Movaghar[22,26], Mor Nahum[23], Torbjörn E M Nordling[27], Thea Norman[9], Robert Penner[25,28,29], Sushmita Roy[6,7], Krishanu Saha[7,22,26], Asif Salim[30], Siamak Sorooshyari[23], Vassili Soumelis[31], Alit Stark-Inbar[23], Audra Sterling[22,32], Gustavo Stolovitzky[10,11], S S Shiju[30], Jing Tang[4,5], Alen Tosenberger[25,33], Thomas Van Vieet[23], Krister Wennerberg[4] & Andrey Zinovyev[15]

[14]The Institute of Mathematical Sciences, HBNI, CIT Campus, Taramani, Chennai, India. [15]Institut Curie, PSL Research University, Mines ParisTech, Inserm U900, Paris, France. [16]Institute of Problems of Mechanical Engineering, Russian Academy of Sciences, St. Petersburg, Russia. [17]Department of Computer Sciences, University of Wisconsin–Madison, Madison, Wisconsin, USA. [18]Department of Social Work, University of Wisconsin–Madison, Madison, Wisconsin, USA. [19]Institute of Health Economics and Health Care Management, Helmholtz Zentrum München (GmbH)–German Research Center for Environmental Health, Neuherberg, Germany. [20]Department of Computer Science, Aalto University, Helsinki, Finland. [21]Helsinki Institute for Information Technology HIIT, Aalto University, Helsinki, Finland. [22]Waisman Center, University of Wisconsin–Madison, Madison, Wisconsin, USA. [23]Posit Science, San Francisco, California, USA. [24]Laboratoire Epigénétique et Cancer, CNRS FRE 3377, CEA Saclay, Gif-sur-Yvette, France. [25]Institut des Hautes Etudes Scientifiques (IHES), Bures-sur-Yvette, France. [26]Department of Biomedical Engineering, University of Wisconsin–Madison, Madison, Wisconsin, USA. [27]Department of Mechanical Engineering, National Cheng Kung University, Tainan, Taiwan. [28]Math and Physics Departments, California Institute of Technology, Pasadena, California, USA. [29]Centre for the Quantum Geometry of Moduli Spaces, Aarhus University, Aarhus, Denmark. [30]Indian Institute of Space Science and Technology, Department of Space, Trivandrum, India. [31]Institut Curie, PSL Research University, Inserm U932, Paris, France. [32]Department of Communication Sciences and Disorders, University of Wisconsin–Madison, Madison, Wisconsin, USA. [33]Unité de Chronobiologie Théorique, Faculté des Sciences, Université Libre de Bruxelles (ULB), Brussels, Belgium.

1. Alipanahi, B., Delong, A., Weirauch, M.T. & Frey, B.J. *Nat. Biotechnol.* **33**, 831–838 (2015).
2. Wang, B. *et al. Nat. Methods* **11**, 333–337 (2014).
3. Kibble, M. *et al. Drug Discov. Today* **21**, 1063–1075 (2016).
4. Eduati, F. *et al. Nat. Biotechnol.* **33**, 933–940 (2015).
5. Sieberts, S.K. *et al. Nat. Commun.* 7, 12460 (2016).
6. Costello, J.C. *et al. Nat. Biotechnol.* **32**, 1202–1212 (2014)