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

Chloé-Agathe Azencott[1,2,3], Tero Aittokallio[4,5], Sushmita Roy[6,7], DREAM Idea Challenge Consortium, Thea Norman[8], Stephen Friend[8], Gustavo Stolovitzky[9,10] and Anna Goldenberg[11,12]

Affiliations

[1]MINES ParisTech, PSL-Research University, CBIO-Centre for Computational Biology, 35 rue St Honoré 77300 Fontainebleau, France.
[2]Institut Curie, 75248 Paris Cedex 05, France.
[3]INSERM, U900, 75248 Paris Cedex 05, France.
[4]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Finland.
[5]Department of Mathematics and Statistics, University of Turku, Finland.
[6]Dept. of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI, USA.
[7]Wisconsin Institute for Discovery, University of Wisconsin-Madison, Madison, WI, USA.
[8]Sage Bionetworks, 1100 Fairview Avenue North, Seattle, WA 98109, USA.
[9]IBM Computational Biology Center, Yorktown Heights, NY 10598, USA.
[10]Department of Genetics and Genomics Sciences, Icahn School of Medicine at Mount Sinai, 1425 Madison Avenue, New York, NY 10029, USA.
[11]Genetics and Genome Biology, SickKids Research Institute, 686 Bay St, Toronto M5G0A4, Canada.
[12]Department of Computer Science, University of Toronto, Toronto, Canada.

DREAM Idea Challenge Consortium

Ankit Agrawal[13], Tero Aittokallio[4,5], Chloé-Agathe Azencott[1,2,3], Emmanuel Barillot[14], Nikolai Bessonov[15], Deborah Chasman[7], Urszula Czerwinska[14], Alireza Fotuhi Siahpirani[16], Stephen Friend[8], Anna Goldenberg[11,12], Jan Greenberg[17], Manuel Huber[18], Samuel Kaski[19,20], Christoph Kurz[18], Marsha Mailick[21], Michael Merzenich[22], Nadya Morozova[23,24], Arezoo Movaghar[21,7,25], Mor Nahum[22], Torbjörn E. M. Nordling[26], Thea Norman[8], Robert Penner[24,27,28], Sushmita Roy[6,7], Krishanu Saha[21,7,25], Asif Salim[29], Siamak Sorooshyari[22], Vassili Soumelis[30], Alit Stark-Inbar[22], Audra Sterling[21,31], Gustavo Stolovitzky[9,10], Shiju SS[29], Jing Tang[4,5], Alen Tosenberger[32,24], Thomas Van Vieet[22], Krister Wennerberg[4] & Andrey Zinovyev[14]

[13]The Institute of Mathematical Sciences, HBNI, CIT Campus, Taramani, Chennai, 600113, India.
[14]Institut Curie, PSL Research University, Mines Paris Tech, Inserm, U900, F-75005, Paris, France.
[15]Institute of Problems of Mechanical Engineering, Russian Academy of Sciences, 199178 St. Petersburg, Russia.
[16]Dept. of Computer Sciences, University of Wisconsin-Madison, Madison, WI, USA.
[17]Department of Social Work, University of Wisconsin-Madison, Madison, WI, USA.
[18]Institute of Health Economics and Health Care Management, Helmholtz Zentrum München (GmbH) - German Research Center for Environmental Health, Ingolstädter Landstr. 1, 85764, Neuherberg, Germany.
[19]Department of Computer Science, Aalto University, Helsinki, Finland.
[20]Helsinki Institute for Information Technology HIIT, Aalto University, Helsinki, Finland.
[21]Waisman Center, University of Wisconsin-Madison, Madison, WI, USA.
[22]Posit Science, 160 Pine Street, San Francisco, CA, USA.
[23]Laboratoire Epigenetique et Cancer, CNRS FRE 3377, CEA Saclay, 91191 Gif-sur-Yvette, France
[24]Institut des Hautes Etudes Scientifiques (IHES), 91440 Bures-sur-Yvette, France.
[25]Department of Biomedical Engineering, University of Wisconsin-Madison, Madison, WI, USA.

[26]Dept. of Mechanical Engineering, National Cheng Kung University, No. 1 University Road, Tainan 70101, Taiwan.
[27]Math and Physics Departments, Caltech, Pasadena, CA 91125 USA.
[28]Centre for the Quantum Geometry of Moduli Spaces, Aarhus University, Denmark.
[29]Indian Institute of Space Science and Technology, Department of Space, Trivandrum, India.
[30]Institut Curie, PSL Research University, Mines Paris Tech, Inserm, U932, F-75005, Paris, France.
[31]Department of Communication Sciences and Disorders, University of Wisconsin-Madison, Madison, WI, USA.
[32]Unite de Chronobiologie Theorique, Faculte des Sciences, Universite Libre de Bruxelles (ULB), Brussels, Belgium.

Over the last two decades we have witnessed an explosion in the amount and diversity of data collected in biological and medical studies. This data is 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 modelling 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 Connectivity Map, to name a few, have unquestionably led to important and actionable findings such as identifying treatment targets[1] or gaining insight into gene regulatory processes. However, such data could have been even more useful. For example, in our own work on glioblastoma subtype discovery[2] , we could only use 46% of the TCGA samples due to missing measurements, reducing the power of the study. In another example, the fixed concentration levels of small-molecule compounds in the Connectivity Map were sub-optimal 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 [4,5,6]. For instance, in a DREAM challenge predicting response to drugs in rheumathoid arthirtis patients, using the largest available collection of SNP data did not improve predictions over clinical predictors[5]. In a toxicogenetic challenge, GWAS data by itself was not predictive but together with RNA-seq available for only 38% of the patients the results were markedly better[4]. Finally, in a DREAM challenge assessing and improving drug sensitivity prediction algorithms, having data of many omics modalities did not provide an advantage over predictors that used gene expression data alone[6]. We concede that these situations could result because some computational models may just be not good enough for the task. However, the fact that none of several dozens of independent expert teams had success in solving the problems using the same data suggests that, alternatively, 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-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 to speed up data collection but also to yield better biological insights, thanks to the exploitation of more appropriate data. Recent successes in physics, such as the

*1*https://cancergenome.nih.gov/researchhighlights/tcgainaction/tcga-data-used-for-loxo101-drug-development

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 where 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.

Another 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 both from bulk populations and single cells. These data will be collected in a model-guided manner, where the initial model is a consensus derived from published datasets to avoid duplication of experimental effort and enable maximal discovery. The resulting data set will serve as a better gold standard to validate computational predictions from existing and new inference methods and help identify the most informative datasets for regulatory network discovery.

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

References:

1. Alipanahi, B., Delong, A., Weirauch, M.T. & Frey, B.J. *Nat. Biotechnol.* **33**, 831–838 (2015).Document7Document7Document7Document7Document7
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. Castello, J.C. et al. *Nat. Biotechnol.* **32**, 1202–1212 (2014)