SciBX
Science-Business eXchange

DREAM team

By Chris Cain, Senior Writer

The DREAM Project has teamed up with **Sage Bionetworks** and announced four open challenges in computational biology that tackle issues relevant to drug discovery and development, including prediction of drug responses. The publication of results from their first collaboration on improving breast cancer prognosis provides a case study for the team's approach.[1,2]

DREAM—Dialogue on Reverse Engineering Assessment and Methods—was founded in 2006 by Gustavo Stolovitzky and Andrea Califano to organize open challenges aimed at improving computational biology. DREAM is sponsored by **Columbia University**, the **NIH**, **IBM Corp.** and **The New York Academy of Sciences**. Past challenges have tackled problems ranging from predicting drug sensitivity of cancer cell lines to modeling epitope-antibody interactions.

> "We believe that the attractor metagenes reflect the underlying biological mechanisms precisely, and we think of them as bioinformatic hallmarks of cancer."
> —Dimitris Anastassiou, Columbia University

Stolovitzky is a manager of functional genomics and systems biology at IBM's computational biology center. Califano is a professor of systems biology and chief of the Division of Biomedical Informatics at Columbia.

DREAM has launched sets of three to five challenges each year for the past six years. Typically, the organization hosts an experimental data set and then asks computational biologists to generate a predictive model that can explain the data. For each challenge, researchers submit computational models to DREAM over a defined time period of a few months. The winning team(s) who develop the best-performing model or models are then invited to an annual conference to discuss the results.

Last year, as part of the DREAM7 series of challenges, Stolovitzky said DREAM changed tack and partnered more closely with supportive organizations than it had in the past. DREAM launched three initiatives with partners—the DREAM Phil Bowen ALS Prediction Prize4Life Challenge to predict disease progression in amyotrophic lateral sclerosis (ALS), the **National Cancer Institute**–DREAM Drug Sensitivity Prediction Challenge to predict the response of cancer cell lines to a set of small molecules and the Sage Bionetworks–DREAM Breast Cancer Prognosis Challenge.

Unlike earlier challenges, the ALS challenge was run in collaboration with open-innovation company **InnoCentive Inc.**, whereas the breast cancer challenge was run on Sage's Synapse online platform. The challenges are now closed, and Stolovitzky said results from the National Cancer Institute and ALS challenges are being prepared for publication.

Stolovitzky said that as the breast cancer challenge progressed, "it was so clear that the vision and outlook for what we were doing was in sync. We had the know-how for operating these challenges, and they had a great software platform and engineering experience, so it was a good marriage."

In particular, he said, Synapse allowed DREAM to achieve a long-standing goal of enabling independent researchers to share data in real time and reproduce each other's results. "The platform allows people to input their own algorithms and have everyone take a look at them. This allowed us to accomplish shared goals such as ensuring the reproducibility of results and of how we score models," he said.

Previously, some interaction among teams could take place through an online discussion forum while the challenges were active. The extent of this communication was limited, though, because no online interface existed for the teams to easily share the experimental methodology developed over the course of the challenge.

Stolovitzky told *SciBX* that the limited ability to share data was due to resource constraints. "It has to do with the history of how DREAM grew organically; it was sort of a garage effort at first. Once the idea got traction, we launched additional challenges but never got a complete set of funding agencies to fully support this outside of limited support for individual challenges. Because of this we didn't have an infrastructure to create a platform to enable collaboration."

On February 19, DREAM announced it was joining with Sage to collaboratively run challenges using the Synapse platform going forward. Last month at the Sage Commons Congress in San Francisco, the first four challenges from the partners were announced as DREAM8 (*see* **Table 1, "Sage Bionetworks–DREAM Project spring 2013 challenges"**).

Although the challenges deal with diverse sources of data, Sage president, cofounder and director Stephen Friend said the framework upon which Synapse is built is adaptable.

"For the breast cancer challenge, we built the necessary tools into the system as the challenge got up and going in real time. For added functionality, such as dealing with proteomic data or imaging data, it requires little additional effort," he said.

Breast cancer pilot

The results from the breast cancer challenge provide a detailed example of how the future Sage-DREAM challenges will operate.[1,2] The goal of the challenge was to take available gene expression, copy number and clinical data and use computational modeling to develop an improved prediction methodology for breast cancer prognosis.

The data were sourced from METABRIC, a large, publically available data set of clinical and genomic information from 1,981 patients with breast cancer.[3]

About a decade ago, Friend participated in the development of a 70-gene prognosis profile for breast cancer that eventually gave rise to **Agendia B.V.**'s marketed MammaPrint prognostic test for breast cancer recurrence.[4]

"Ten years ago, we developed this method using breast cancer data, but the methodology hadn't really evolved from there. So we asked if the crowd could evolve a better variation of the approach," he said.

The METABRIC data were adapted into Synapse, and 354 participants

Table 1. Sage Bionetworks–DREAM Project spring 2013 challenges. Sage Bionetworks and The DREAM Project have announced four open-innovation computational challenges that tackle issues relevant to drug discovery and development. The challenges will run between May and September. The partners expect to announce another round of challenges in the fall.
Source: Sage Bionetworks

Title	Description	Data source	Sponsor
HPN-DREAM Breast Cancer Network Interference Challenge	Use quantitative proteomic data to: (i) build network models that represent the active pathways and their response to different stimuli during drug treatment; (ii) predict the responses of phosphoproteins to various drugs; and (iii) propose new visualization strategies for the high-dimensional data sets	**Oregon Health & Science University; The University of Texas MD Anderson Cancer Center; The Netherlands Cancer Institute**	**Heritage Provider Network Inc.** (HPN); **National Cancer Institute**
NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge	Use genetic and toxicology data to build computational models that can predict: (i) the toxic response of individuals to each chemical based on genetics and genomics data; and (ii) the parameters of distribution for the toxic effects of each chemical based primarily on chemical information about the compounds being evaluated	**National Institute of Environmental Health Sciences** (NIEHS); **National Center for Advancing Translational Sciences** (NCATS); **The University of North Carolina at Chapel Hill** (UNC)	To be announced
National Brain Tumor Society–DREAM Cancer Prediction Challenge	Determine whether systems biology–based models of human glioblastoma multiforme (GBM) are sufficiently advanced to allow the correct prediction of single agents or combinations of drugs that may abrogate tumorigenesis or significantly delay tumor growth *in vivo*	Multiple sources to be announced	**National Brain Tumor Society**
Whole-Cell Parameter Estimation DREAM Challenge	Refine a whole-cell computational model describing the biology of *Mycobacterium genitalium* by predicting a subset of the kinetic parameters used to represent fundamental biological processes, with the goal of determining how accurately the kinetics of cellular processes can be reverse engineered	**Stanford University**	To be announced

registered for the challenge to analyze the data and develop prognostic models. The source code for each model was made available on Synapse to encourage collaboration between participants. To promote competition and model improvement, results from the challenge were updated in a real-time online leaderboard that had not been available in earlier DREAM challenges.

After two phases of model development and validation on subsets of the METABRIC data, the final models were tested on an independent data set from 184 patients to determine a winner.

The winning model came from a research team at Columbia University led by Dimitris Anastassiou, a professor of electrical engineering. The model built upon his group's previous work defining attractor metagenes, which are signatures of coexpressed genes in cancer identified through an iterative computational approach.[5]

This approach identified sets of coexpressed genes associated with particular cancer phenotypes, including mitotic chromosome instability, mesenchymal transition and lymphocyte-specific immune recruitment.

"We believe that the attractor metagenes reflect the underlying biological mechanisms precisely, and we think of them as bioinformatic hallmarks of cancer," said Anastassiou.

Models were scored based on their prediction of the concordance index (CI). For every two randomly selected patients, the CI is the probability that a model will correctly predict which of the two patients will die before the other. So for random chance, the CI would be 0.5. On the test set, the attractor metagene signature had a CI of 0.756, whereas the previously identified 70-gene signature had a CI of 0.60.

Results were published in *Science Translational Medicine*, which embedded peer reviewers into the challenge process to evaluate the results and help determine criteria for selecting a winner.

Crowd to clinic

Agendia CMO Neil Barth told *SciBX* that the data-analysis crowdsourcing approach could provide a future effective model for diagnostics development.

"This provides a unique platform for the development and refinement of models. As we are beginning to enrich databases with all kinds of new information at all levels, including gene and protein expression, this kind of modeling to get to clinical answers is probably the most efficient way to go about it," he said.

He cautioned that several steps need to be taken to further clinically validate the results. First and foremost, Barth said, it would be important to set and validate a threshold at which the test could provide clinically meaningful results. "If you have a test like MammaPrint, you have to set a threshold of minimal performance for the low-risk population. So, for example, in our case, low risk means having a 10-year survival rate of 90% or better. There is no threshold set for the performance of these models; it's simply which one is performing statistically better, not necessarily held against the defined threshold of outcome," said Barth. "These models were designed to try to get to the best *p* value, but you aren't given that luxury in the clinical arena."

He also noted that two other differences between MammaPrint and this approach are that the marketed test is focused on the risk of an individual patient, not a cohort of patients, and that the test does not look at metagenetic signatures.

The new models blend both clinical data and gene expression data, which Barth said realistically resembles how most doctors make decisions using marketed gene expression tests. However, he noted that diagnostic development is different and more difficult because an expression-based test must show a statistically significant benefit by itself as a stand-alone assay.

Anastassiou added that commercially available tests including Oncotype DX and MammaPrint use genes related to the attractor metagenes, and he plans to test whether replacing any of the genes in these tests could improve the accuracy of the products.

Oncotype DX is a breast cancer prognostic marketed by **Genomic Health Inc.**, which declined to comment.

Barth also said that once approved, a test such as MammaPrint cannot be significantly changed without requiring further clinical validation. “If we had the ability to have an open-source community look at the signature, there is no question in my mind we could be further ahead at bringing to the clinic a more optimized tool. I have no doubt this type of approach has the opportunity to collaboratively improve these signatures.”

Anastassiou said Columbia has filed patent applications covering biomarkers used in its model. He did say open sharing in community challenges such as this is vital for developing better diagnostics.

Stolovitzky agreed. “I work for IBM, and I clearly understand the value of IP, but at the same time I believe that in some ways being too concerned with privacy and IP protection can delay progress,” he said. “That doesn’t mean that people should not consider filing a patent on their methods, but once you are a part of a challenge, our thinking is that you should share with others. If you are the best performer, let us know what you did, allow us to see that there is reproducibility and work with us to advance the field.”

He added, “I won’t claim this is completely sorted out, but I think there is a place for IP and a place for collaborative learning, and we are trying to sort out the right way to do this.”

Cain, C. *SciBX* 6(18); doi:10.1038/scibx.2013.430
Published online May 9, 2013

REFERENCES

1. Margolin, A.A. *et al. Sci. Transl. Med.*; published online April 17, 2013; doi:10.1126/scitranslmed.3006112
   Contact: Stephen H. Friend, Sage Bionetworks, Seattle, Wash.
   e-mail: friend@sagebase.org
   Contact: Adam A. Margolin, same affiliation as above
   e-mail: margolin@sagebase.org
2. Cheng, W.-Y. *et al. Sci. Transl. Med.*; published online April 17, 2013; doi:10.1126/scitranslmed.3005974
   Contact: Dimitris Anastassiou, Columbia University, New York, N.Y.
   e-mail: da8@columbia.edu
3. Curtis, C. *et al. Nature* **486**, 346–352 (2012)
4. Van de Vijver, M.J. *et al. N. Engl. J. Med.* **347**, 1999–2009 (2002)
5. Cheng, W.-Y. *et al. PLoS Comput. Biol.* **9**, e1002920; published online Feb. 21, 2013; doi:10.1371/journal.pcbi.1002920

COMPANIES AND INSTITUTIONS MENTIONED

Agendia B.V., Amsterdam, the Netherlands
Columbia University, New York, N.Y.
The DREAM Project, Seattle, Wash.
Genomic Health Inc. (NASDAQ:GHDX), Redwood City, Calif.
IBM Corp. (NYSE:IBM), Armonk, N.Y.
InnoCentive Inc., Waltham, Mass.
National Cancer Institute, Bethesda, Md.
National Institutes of Health, Bethesda, Md.
The New York Academy of Sciences, New York, N.Y.
Sage Bionetworks, Seattle, Wash.