THE AUTHOR FILE

# Gustavo Stolovitzky

## Boosting accuracy by building consensus

Gustavo Stolovitzky does not trust his results readily. A computational biologist at the IBM T.J. Watson Research Center, Stolovitzky cranks through genomic data in search of circuits that control cell behavior. His algorithms sift for patterns and then infer components of a gene regulatory network. But such projects can stumble quickly into the 'self-assessment trap', he says. "It is very easy to fool oneself into thinking that you have predictions that are true, and it is difficult to validate those predictions."

Gustavo Stolovitzky

Prediction methods for regulatory circuits may produce desired results based on researchers' input rather than patterns intrinsic to data. Without validated gold standards or ground truth, it is hard to tell whether development of a researcher's computational tool has been self-correcting or self-fulfilling. People in the field have a saying, explains Stolovitzky: "When you torture a data set, it will confess whatever you want."

Stolovitzky wondered whether systems biology could benefit from its own version of Critical Assessment of Structure Prediction (CASP). In this biannual collaborative experiment (organizers eschew terms such as 'competition'), teams of researchers predict folding for a protein whose structure has been solved but not yet published. After making some inquiries, Stolovitzky and other scientists established the Dialogue on Reverse Engineering Assessment and Methods (DREAM) project, and the first set of challenges was launched in 2007. This year's DREAM7 challenges, still open to participants, focus on cancer and amyotrophic lateral sclerosis.

When writing up the results of DREAM2 in 2008, Stolovitzky began wondering what would happen if he aggregated the predictions made by multiple regulatory network inference methods. Most likely, he thought, "the not-best prediction would degrade the best prediction." But in preliminary tests of this idea, Stolovitzky found that the aggregate prediction actually outperformed the best independent prediction. "It's like the positives summed up and the negatives didn't subtract," he says.

Stolovitzky says he became 'obsessed' with figuring out whether aggregating different methods could be a reliable way to improve performance. He recalls a 3 a.m. insight on the underlying mathematics. "I said 'I understand. This is not a fluke; this is something that has to be. There is a robustness to the aggregation process.'"

The DREAM5 gene regulatory network inference challenge was designed in part to test this idea systematically. The outcome, published in this issue of *Nature Methods*, is that an aggregate method may not always be the best, but it will be among the best. Stolovitzky and colleagues also describe techniques for aggregating network inference methods and provide a web interface to construct consensus networks from multiple methods.

In the DREAM5 challenge, teams inferred genome-wide transcriptional regulatory networks for microarray data from *Saccharomyces cerevisiae*, *Escherichia coli* and *Staphylococcus aureus* as well as for simulated data. Stolovitzky and his coauthors assessed 35 distinct methods: 29 submitted by participants plus 6 well-known 'off-the-shelf' methods. *S. aureus* lacked sufficient data to create a gold-standard network, and the aggregated results provide the first genome-wide prediction of its gene regulatory network. For the other data sets, the authors evaluated predictions for individual methods and aggregates.

When assessed individually, the strengths and weaknesses of various methods clustered by the underlying approach. For example, regression and Bayesian strategies worked best for linear gene regulatory cascades, but mutual-information and correlation-based methods worked best for feed-forward loops, in which transcription factors act jointly rather than sequentially. The best performers for *E. coli, S. cerevisiae* and simulated data all came from different inference approaches. "There is no 'one-size-fits-all' algorithm for gene network inference," says Stolovitzky. But integration still improved performance, particularly when disparate methods were assembled together.

"It's like the positives summed up and the negatives didn't subtract."

Other advances come when participants meet together to look at how methods performed, says Stolovitzky. Normally, researchers from different teams will not be well versed in each other's data, thus limiting discussion. Not so with DREAM participants. "They can really have a lingua franca for how the methods worked because they are applying them to the same data." To keep discussions going, Stolovitzky is careful that no one is embarrassed for participating. Only the best performers are named, and he studiously avoids words such as 'winners' and 'losers'.

Results are better, he says, when there are as many participants as possible. "More divergent ideas, if partially right, will improve the predictions."

**Monya Baker**

Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nat. Methods* **9**, 799–807 (2012).