Genomic data are growing, but what do we really know?

March 20 2013

"We live in the post-genomic era, when DNA sequence data is growing exponentially", says Miami University (Ohio) computational biologist Iddo Friedberg. "But for most of the genes that we identify, we have no idea of their biological functions. They are like words in a foreign language, waiting to be deciphered." Understanding the function of genes is a problem that has emerged at the forefront of molecular biology. Many groups develop and employ sophisticated algorithms to decipher these "words". However, until now there was no comprehensive picture of how well these methods perform, "To use the information in our genes to our advantage, we first need to take stock of how well we are doing in interpreting these data".

To do so, Friedberg and his colleagues, Predrag Radivojac, of Indiana University, Bloomington IN and Sean Mooney, Buck Institute for Research on Aging, Novato CA organized the Critical Assessment of protein Function Annotation, or CAFA. CAFA is a community-wide experiment to assess the performance of the many methods used today to predict the functions of proteins, the workhorses of the cell coded by our genes.

Thirty research groups comprising 102 scientists and students participated in CAFA, presented a total of 54 methods. The participating groups came from leading universities in North America, Europe, Asia and Australia. The groups participated in blind-test experiments in which they predicted the function of protein sequences for which the functions are already known but haven't yet been made publicly available.

Independent assessors then judged their performance.

The results are published in this month's issue of *Nature Methods* co-authored by members of all the participating groups, with Friedberg and Radivojac as lead authors. Fifteen companion papers have been published in a special issue of *BMC Bioinformatics* detailing the methods.

"We have discovered a great enthusiasm and community spirit", said Friedberg, who since 2005 has been organizing Automated Function Prediction (AFP) meetings internationally. This, despite the competitive environment in which research groups want their methods to perform better than their peers' methods. Overall, throughout CAFA there was a highly collegial spirit, and a willingness to share information and science. "Everyone recognized that this is an important endeavor, and that only by a group effort can we move the field forward and learn to harness the deluge of genomic data, turning it into useful information."

"For the first time we have broad insight into what works, where improvement is needed, and how we should move the field forward. We will continue running CAFA in the future, as we are confident it will only help generate better methods to understand the information locked in our genomes, and those of other organisms," Friedberg said.

The initial analysis suggests that algorithms combining disparate prediction clues taken from different knowledge-bases provide more accurate predictions. The lead methods combined data from phylogenetic, gene-expression and protein-protein interaction data to provide predictions.

More information: Radivojac et al, *Nature Methods*: www.nature.com/nmeth/journal/v … full/nmeth.2340.html
BMC Bioinformatics companion papers:

www.biomedcentral.com/bmcbioin … cs/supplements/14/S3
The Automated Function Prediction Special Interest Group web site: biofunctionprediciton.org

Provided by Miami University