FUNCTIONAL GENOMICS

Learning to prioritize

The number of species with sequenced genomes is rising rapidly, but annotating the protein functions encoded in these genomes from scratch would be an expensive and time-consuming process. Instead, as a recent study shows, existing data from model organisms can be used to design a streamlined set of experiments to annotate the genomes of related species.

Many large-scale expression microarray experiments have been done for the model yeast *Saccharomyces cerevisiae*. Guan and colleagues used these data to prioritize experiments for functionally annotating the genome of a related species, *Saccharomyces bayanus*. Their approach was based on the fact that microarray studies generally provide information about genes involved in multiple biological processes, often including those that are not directly related to the original biological question being asked.

The first step was, on the basis of *S. cerevisiae* expression profiles, to rank treatment conditions in terms of how informative they are about the genes involved in a particular process. This was done using a

computational method to determine the power of the data sets to make predictions for each process. Next, algorithms were used to assess redundancy in the information provided under each condition. From these analyses, Guan and colleagues predicted that a small subset of the thousands of array experiments that have been done in *S. cerevisiae* — just 10% of the arrays — would be sufficient for a first-pass functional annotation of the *S. bayanus* genome.

To validate their approach, the authors did the suggested set of *S. bayanus* microarray experiments and confirmed that they have predictive power for a wide range of biological processes. Furthermore, the accuracy of these predictions was confirmed: a subset of *S. bayanus* genes have functionally annotated orthologues in *S. cerevisiae*, and the predictions made from the array data about these genes matched up well with those made on the basis of orthology. The authors then used their approach to select a second tier of experiments for the smaller set of functions that were not captured by the initial arrays.

Guan and colleagues have made their approach for yeast gene expression data freely available to others (see website below). Similar methods should be applicable to other types of genome-wide data sets and, as the authors point out, *S. bayanus* and *S. cerevisiae* are separated by 20 million years of evolutionary history, showing that this approach works even for not-so-close relatives.

Louisa Flintoft

ORIGINAL RESEARCH PAPER Guan, Y., Dunham, M., Caudy, A. & Troyanskaya, O. Systematic planning of genome-scale experiments in poorly studied species. *PLoS Comput. Biol.* **6**, e1000698 (2010)

WEBSITE

Functional Coverage Analysis of Microarray Data and Experiment Recommender for Yeast: <http://exp recommender.princeton.edu>