SYSTEMS BIOLOGY

A next-gen ontology

© 2013 Nature America, Inc. All rights reserved.

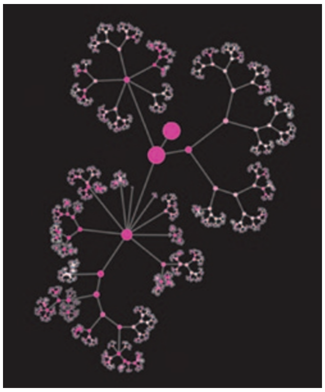
An approach to cluster and organize systems biology data yields NeXO, a data-driven ontology.

Where biology was small before, it now is frequently large. Data points have become data sets. Research studies now commonly examine not just one gene or protein but all of them. It has become a truism to say that this situation produces formidable challenges when it comes to analysis of the resulting vast data sets.

The Gene Ontology (GO) has proven to be a widely used resource for analyzing systems biology data. Built manually from the literature by expert curators, GO attempts to synthesize available biological knowledge into an organized hierarchy of biological terms, linking genes to function. Large-scale data sets are frequently analyzed in light of GO to assess whether they capture meaningful patterns of gene function.

In a recently published study, Trey Ideker and Janusz Dutkowski at the University of California, San Diego, and their colleagues, effectively turn this way of thinking on its head. They explore, in other words, whether 'omics' data sets themselves have meaningful ontological structure, akin to that of GO, and whether methods can be found to reveal it. The result is a parallel ontology to GO, which they call the network-extracted ontology, or NeXO.

Ideker and colleagues begin with four large yeast data sets, describing protein-protein interactions, genetic interactions, gene coexpression and an integrated functional network that includes several data types. They integrate these data sets into a single network and then chip away at it with mathematical tools, searching for its underlying structure. First, they cluster sets of genes according to their patterns of interactions, trying to maximize the stability of the resulting nodes. Then they transform the structure of the resulting binary tree into one that more accurately reflects the branched, nonbinary nature of biological relationships. Finally, they align the resulting tree with GO, assigning GO terms to NeXO where appropriate and identifying discrepancies between the two.



The yeast NeXO: a network-extracted ontology from yeast data. Reprinted from *Nature Biotechnology*.

The structure that emerges—including about 4,000 terms and close to 6,000 relationships among them—clearly reflects existing biological knowledge, although it is based purely on patterns of interactions between genes in the starting data sets. NeXO has major mitochondrial, membrane and intracellular branches; these in turn branch further into subparts that correspond to recognizable cellular components (the proteasome or the cytoskeleton, for example) and reflect previous knowledge about their internal organization.

About one-third of NeXO terms map to GO, and NeXO captures about two-thirds of terms in the cellular components arm of GO. (The present version of NeXO captures fewer terms, about a quarter each, from the biological processes and molecular functions arms of GO.) NeXO performs as well as or better than GO in a functional enrichment analysis of two independent data sets that were not used in its construction.

As a consequence of the way it is extracted—computationally, from large data sets—NeXO sidesteps one of the limitations of the manually curated GO, which inevitably focuses on well-studied biological entities. By contrast, NeXO includes many terms that correspond to biological entities without names, descriptions or functional annotations and can thus be used to generate and test hypotheses about previously uncharacterized components. As an illustration of this, Ideker and colleagues study 73 genes within a NeXO term associated with the Golgi apparatus, of which almost half were previously uncharacterized. They determine quantitative genetic interaction profiles between these genes and thus experimentally confirm relationships suggested by the NeXO.

Another use for this ontology, Ideker and colleagues propose, is as a tool for GO curators. Terms in NeXO that have not yet been incorporated into GO but that have strong support in the network data can be prioritized for manual curation. Indeed, the researchers work together with GO editors to refine and modify GO annotations.

NeXO is thus not only a demonstration that there are biologically meaningful ontological hierarchies embedded in omics data sets, valuable though this observation is, but a tool for further investigation. The methods used to generate NeXO could also be used to assess new data sets: does a new data set capture the same underlying structure, or does it suggest differences? Does it add new knowledge or simply reinforce what is already known? A challenge for the future is to address how additional data will be incorporated into NeXO and used to refine existing terms and relationships or to add new ones.

Natalie de Souza

RESEARCH PAPERS
Dutkowski, J. *et al.* A gene ontology inferred from molecular networks. *Nat. Biotechnol.* **31**, 38–45 (2013).