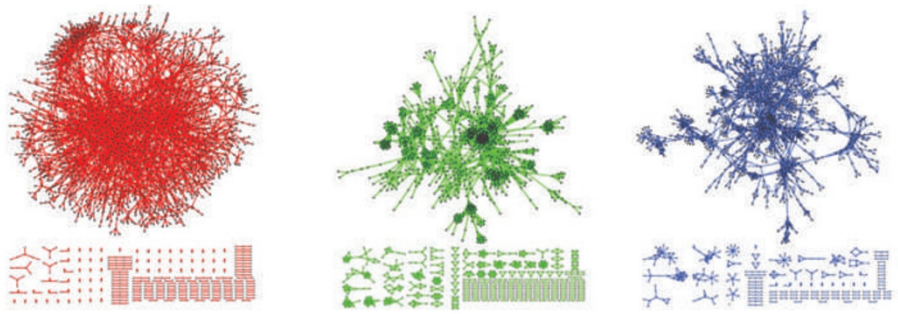
M. VIDAL



Different approaches for identifying protein–protein interactions often reveal unique information.

says Bryant, describing GeneGo's MetaCore software. Being able to overlay a variety of different experimental data from different sources requires careful database curation, she says. At the moment, GeneGo employs 50 scientists to manually mine and curate published literature for studies on protein interaction, gene expression, metabolism and drugs to expand and update its internal database, which now contains more than 120,000 multi-step interaction pathways, each averaging 11 steps, with information on direction, mechanism and feedback along the pathways, along with direct links to literature evidence.

Literature mining is important for building larger interaction databases, but Bryant says it can be especially difficult if the experimental descriptions underlying the results have not been published. Another problem, according to Vidal, is that researchers sometimes have "sociological" biases in terms of which proteins and interactions they will work on and report. "We have learned a lot about the rules of how macromolecules interact, but when you ask how much of the network we have, or what the size of the interactome of a particular species is, if you only used the literature it would be tough to answer those questions," he says.

Tyers is involved with the publicly funded BioGRID (Biological General Repository for Interaction Datasets) initiative, an internationally curated database of molecular interactions. Three years ago, there was an effort to back-curate all the yeast literature for protein and genetic interactions, but now the database contains protein-interaction data from yeast, worms, flies, plants and even humans along with some genetic-interaction data as well. For Tyers, the goal is to accurately mirror the primary literature and distil it into a format that can be used in network biology. "We make no judgement calls on the method or even, within reason, the quality of the data themselves," he says, giving researchers the opportunity to

PLAYING BY THE RULES

When researchers at Plectix BioSystems in Somerville, Massachusetts, began to use their new Cellucidate software to model the epidermal growth factor receptor pathway, they calculated that there were 10^{33} potential states — including all protein complexes and phosphorylation states — for the system. "This is the kind of complexity that scientists have to grapple with when it comes to cell-signalling networks," says Gordon Webster, vice-president of biology at Plectix.

Although not all these potential states necessarily occur in that pathway, when it comes to creating more manageable models for understanding cell signalling researchers face a difficult question: what interaction data do they use in their models? Although many commercial and public databases still rely heavily on the small-scale protein–protein-interaction studies that appear in peer-reviewed literature, the emergence of high-throughput experimental approaches that generate very large interaction data sets is creating the need for a new set of rules.

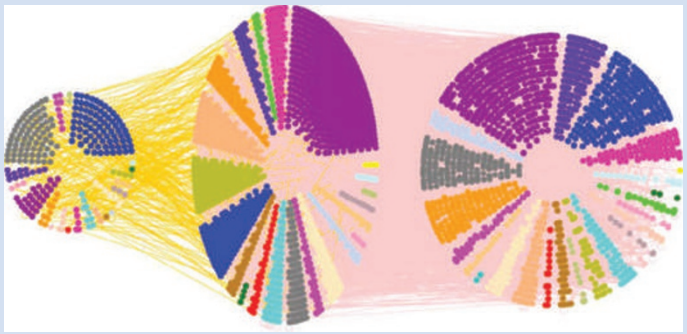
"In practice, what comes out of these high-throughput studies is not a yes/no thing — 'these interact, and these don't' — but in fact they generate a list of interactions and associated probabilities," says Jack Greenblatt from the University of Toronto in Canada. To generate such probabilities for his mass-spectrometry studies, Greenblatt applied a 'gold standard' for protein interactions — a set of protein complexes or interactions in which there is a strong amount of confidence according to the literature — as well as a set of proteins not known to interact with one another as a negative standard. He then tackled the question of whether or not data sets generated by mass spectrometry stacked up against protein-interaction reports seen in peer-reviewed literature.

"What we did in the end was to use the same gold standard to look at the molecular-biology literature," says Greenblatt. After adjusting the cut-off point so that the average confidence score from a high-throughput study matched the confidence score of interactions reported in the literature, he says the interaction data from such studies are no better or worse than what is in the literature.

Marc Vidal, a geneticist at the Dana–Farber Cancer Institute in Boston, Massachusetts, wants to see a similar approach taken with yeast two-hybrid and other binary screens. "Let's roll up our sleeves and decide on a positive and negative gold standard," he says. "But let's also use orthogonal assays to give confidence scores to the interactions."

In January, Vidal and his colleagues published a series of papers[6–9] suggesting the use of new binary interaction assays to build confidence in basic networks produced using yeast two-hybrid data sets. "You say 'OK, this is basic network' and then push that into a framework where all interactions are going to be tested by two or three orthogonal assays. And not only that, but do that under conditions where you have a positive and negative gold standard," says Vidal, adding that the high-scoring interactions can then serve as hypotheses for researchers to test.

M. TYERS



Graphical representation of the current budding-yeast interaction network.

Whether or not these efforts and standards will lead researchers to rely more on large-scale data sets and mine them more deeply will only be known in time. For some, even with confidence measures, large-scale data sets lack information often found in smaller studies. "This is one of the paradoxes that we find when people talk about systems biology. With technology it is very easy to generate spreadsheets of interaction data, but that alone does not represent any knowledge," says Webster.

But for Greenblatt and others, large-scale data sets represent a starting point for further research efforts. "To me, high-throughput studies are just like the conventional literature," he says, "providing a gold mine for people to dig into." N.B.