Untangling the protein web

Researchers have identified thousands of macromolecular interactions within cells. But, as Nathan Blow finds out, joining them up in networks and figuring out how they work still poses a big challenge.

In the spring of 2006, Andrew Emili and Jack Greenblatt from the University of Toronto in Canada and their colleagues published a survey[1] of the global landscape of protein complexes within the yeast *Saccharomyces cerevisiae* in *Nature*. In the same issue, another group of researchers from the drug research company Cellzome in Heidelberg, Germany, also reported[2] on *Saccharomyces* protein complexes. "Those two data sets overlapped nicely, but by no means perfectly," says Mike Tyers, a systems biologist at the University of Edinburgh, UK. "And yet it was essentially the same method and same organism."

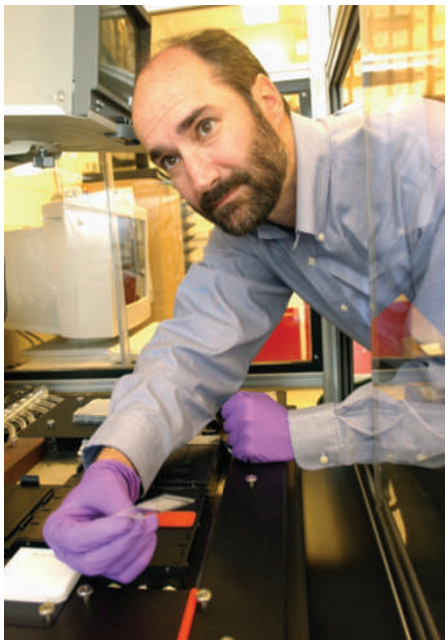
Greenblatt thinks that the two studies highlight something important that is emerging from the current crop of large-scale protein–protein interaction studies. "If you combine data sets you have more information than from any one study alone," he says. This is not to say that one such study is right and the other is wrong: scientists suspect it is more likely that one study often compensates for another's false negatives, revealing true protein interactions that can be missed during a single screen.

"I think the interaction space is very large. Part of the issue is that there is a large range of interaction affinities, and as you start to get down into the weaker interactions those are tougher to detect," says Tyers. He adds that identifying such "moving targets" is not like sequencing DNA, which can be argued to be a more stable target for researchers to aim at.

But the jigsaw pieces are starting to pile up as researchers generate more and more genetic, metabolic and protein-interaction data sets using a diverse array of technologies. This work has been aided in recent years by a number of improved methods and techniques. Add to this recent refinements in computational tools for modelling signalling pathways and it's clear that scientists might be on the cusp of changing the way they look at signalling and information flow in cells.

Embracing diversity

"I think genetic information lays out the blueprints, whereas proteomics is much closer to what is going on in the cell, a molecular manifestation of a phenotype," says Mike Snyder, a biologist at Yale University. When it comes to cataloguing proteins and their interactions, researchers are learning to embrace experimental diversity. "Every approach will usually give an overlapping but distinct set of information," says Snyder. "They all have their strengths and weaknesses."



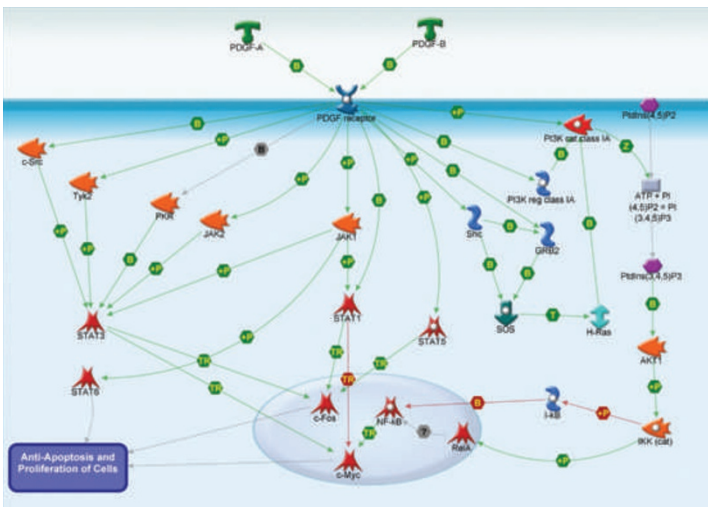
M. SNYDER

Mike Snyder has used protein arrays to explore the yeast interactome.

Tyers and Greenblatt are in a growing group of investigators who are advancing the use of affinity-purification chromatography followed by mass spectrometry to uncover protein interactions in different cell types. In this approach, a protein of interest is tagged with a label that can be used for affinity purification. Although some scientists suspect weaker-interacting protein pairs or transient interactions could be lost during purification, Greenblatt — whose lab relies on tandem affinity purification tags in their purifications — says this is where the use of mass spectrometry helps out. "Mass spectrometry is very sensitive, so even if you lose 90% of the interactor during the affinity purification you can still detect the 10% that is left," he says.

As with the technologies behind protein–protein analysis, researchers are finding that no single labelling tag may be enough to isolate all proteins. Tyers's group recently reinterrogated a section of the yeast proteome using three different tags, each with different properties. "For a number of baits we queried, it made a difference what tag was on it," he says. "Tags can certainly affect the recovery of interactions, consistent with the well-known genetic effects often caused by different tags."

Once a specific protein or protein complex is purified, it is analysed with mass spectrometry. Electro-spray ionization or matrix-assisted laser desorption ionization (MALDI) volatilizes and ionizes peptides, which are analysed on orthogonal or quadrupole time-of-flight (Q-TOF) instruments to identify ions with high mass-to-charge ratio values. Here researchers have benefited greatly from advances by instrument developers. During the American Society for Mass Spectrometry annual conference in Philadelphia, Pennsylvania, in June, Bruker Daltonics of Billerica, Massachusetts, announced its new ultrafleXtreme MALDI TOF/TOF system, and Thermo Fisher Scientific of Waltham, Massachusetts, introduced the LTQ Velos and LTQ Orbitrap Velos devices. Alongside other hardware, such as the Xevo Q-TOF from Waters in Milford, Massachusetts, and the 6500 series of Q-TOF instruments from Agilent Technologies in Santa Clara, California, these machines have improved both the dynamic range and sensitivity of mass analysis; in many cases they also feature integrated upstream separation technologies and improved databases, all of which is making it easier to define a sample's protein composition. For additional detail in the analysis, protein complexes can also be analysed with tandem mass spec-

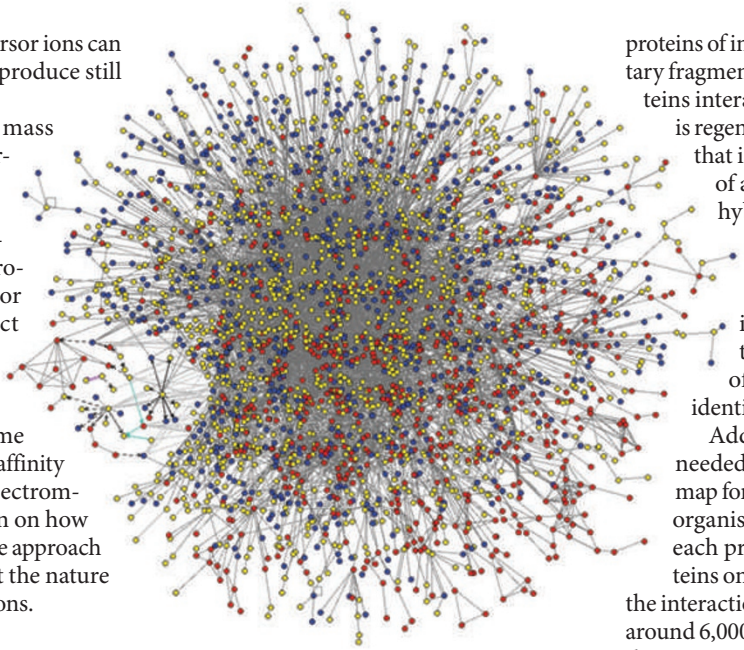


GENEGO

Pathway maps illustrate the complexity of cellular interactions.

trometry, in which selected precursor ions can be smashed into one another to produce still smaller fragments for analysis.

"The big advantage of using mass spectrometry is that it can be performed in a physiological context," says Tyers. Unlike other methods for surveying protein–protein interactions, mass spectrometry can be done on cell lines or even tissue samples, so indirect interactions that depend on more than two proteins or on post-translational protein modifications can be uncovered. Still, some researchers suggest that although affinity purification followed by mass spectrometry gives important information on how proteins interact in complexes, the approach does not reveal everything about the nature and mechanics of those interactions.



M. VIDAL

***Caenorhabditis elegans* interactome map, showing 5,500 protein interactions among 3,000 proteins.**

Yeast shows the way

Binary approaches, such as the yeast two-hybrid assay, can provide different protein interaction information, according to Marc Vidal, a geneticist at the Dana–Farber Cancer Institute in Boston, Massachusetts. Vidal uses the analogy of two football teams facing each other with referees in the middle of the field to explain the differences between the techniques. "The pull-down mass-spectrometry approach will show you the players, referees and field, but not who is passing to whom and in what direction the ball is travelling," he says. "This is where a binary approach comes in."

The yeast two-hybrid assay is arguably the best-known binary approach. It relies on a split transcription factor in which one portion is placed on each of the two proteins being tested for interaction. If the proteins interact, the transcription factor will be regenerated and a reporter gene transcribed, providing a read-out. The assay allows for more the testing of dynamics of protein–protein interactions, such as dissociation rates. "Physical interactions are not everything: you need both edges and arrows to know the dissociation rates as well as other logical aspects of the relationships. Pull-down mass spectrometry is a little short when it comes to those interactions," says Vidal.

The other advantage of the yeast two-hybrid approach is that it presents a more high-throughput solution to studying protein interactions. "The two-hybrid approach is reasonably high-throughput," says Snyder, noting that with robotics a large number of proteins can be tested for potential interactions in a two-by-two format.

Other approaches have also been rising to the surface. "From the probing that we have done, we have picked up interactions that you definitely do not see with other methods," says Snyder of his experience using protein microarrays to explore protein–protein interactions. Protein arrays, which are sold by a number of companies including Invitrogen in Carlsbad, California, RayBiotech in Norcross, Georgia, and R&D Systems in Minneapolis, Minnesota, have not been used as often for large-scale protein-interaction studies as either mass spectrometry or the binary-interaction approaches. "They have had impact in certain areas. Part of the problem is that they have been somewhat expensive, which might be the reason that they have not caught on as much for large-scale studies," says Snyder. Given the potential of protein microarrays to identify unique interactions, he hopes that costs will fall, which could increase their use in large-scale interaction studies.

An orthogonal approach to the yeast two-hybrid assay for detecting protein–protein interactions is the protein-fragment complementation assay (PCA), in which two proteins of interest are attached to complementary fragments of a reporter protein. If the proteins interact with one another the reporter is regenerated providing a direct read-out that is not dependent on transcription of another gene as in the yeast-two-hybrid assay. Steven Michnick and his colleagues used the PCA approach last year[3] to explore the yeast-protein interactome, identifying nearly 2,800 interactions among 1,124 proteins, many of which had not previously been identified by other approaches.

Additional work and tools could be needed to define a complete interaction map for even the most well-characterized organisms. Snyder suspects that in yeast each protein 'sees' about five other proteins on average. But at the moment all of the interactions identified for yeast, which has around 6,000 proteins, add up to far fewer than the potentially 30,000 predicted. "So, there is still a way to go," he says.

Clear pathways

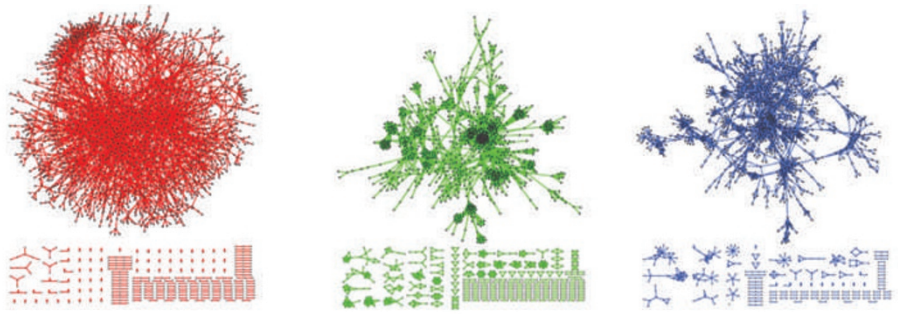
Finding which macromolecules interact is only the first step to figuring out signalling pathways. Researchers also need methods to assemble those interactions into cellular networks, which is where bioinformatics enters the picture. "It is like building a bicycle — you have the wheels, a seat and handlebars, but we provide the steps to put the parts together," says Julie Bryant, vice-president of business development at GeneGo in St Joseph, Michigan, a company specializing in the development of software for cell-signalling and metabolic analysis. GeneGo is not alone here: a growing number of developers are creating tools for the analysis of signalling networks — from those that build model networks based on existing data to systems that use data sets and models to make predictions about the activity of different signalling networks.

"We can take in any kind of experimental data — genomic, proteomic, metabolomic — and overlay them on cell-signalling pathways,"

THERMO FISHER SCIENTIFIC

Advances in mass spectrometry technology are benefiting protein–protein interaction studies.

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.

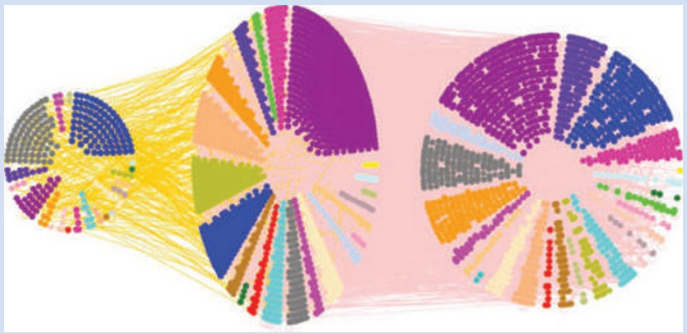
"In practice, what comes out of these high-throughput studies is not a yes/no thing — 'these interact, and these don't' — but in fact they generate a list of interactions and associated probabilities," says Jack Greenblatt from the University of Toronto in Canada. To generate such probabilities for his 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.

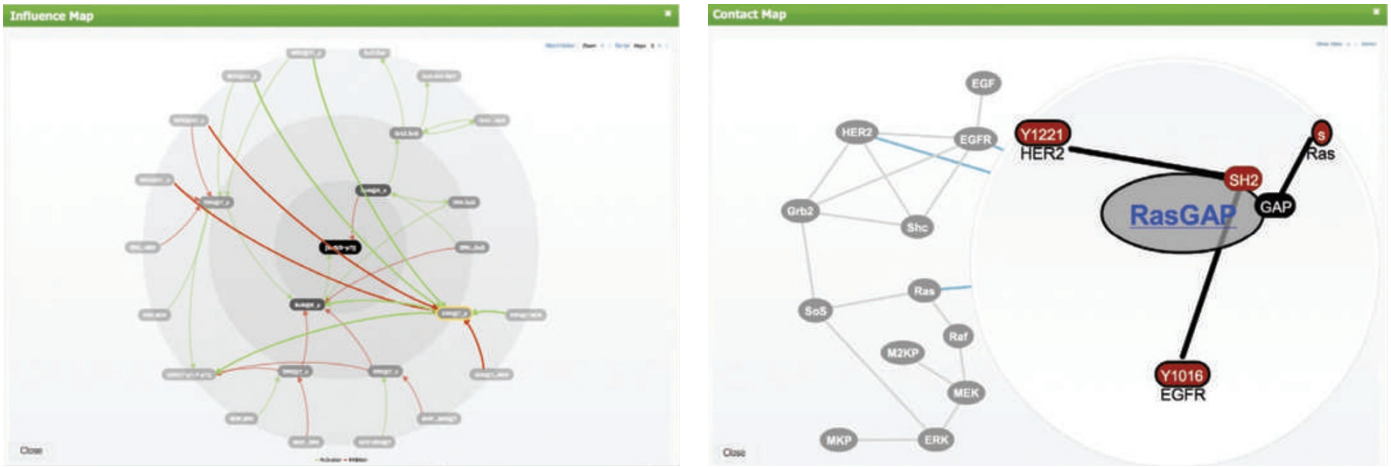
M. TYERS



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.



PLECTIX BIOSYSTEMS

Cell-signalling software packages allow researchers to model and test cellular interaction networks.

extract the maximum amount of information.

A different angle in modelling signalling networks was recently described by Walter Fontana from Harvard University and his colleagues[4]. It uses sets of rules to define relationships between cellular components instead of the more conventional method of defining specific interactions and species using differential equations. Fontana co-founded a company called Plectix BioSystems in Somerville, Massachusetts, which has employed this approach in a web-based system called Cellucidate.

"The system is represented at a very granular level where the participants are allowed to do *in silico* what they would do in real life," says Paul Edwards, chief executive at Plectix. Imagine the city-building computer game SimCity reworked for complex cellular networks, but here the agents of the cell — proteins and other molecules — are the automata instead of colourful animated people. "In that way the model mirrors the behaviour of the living system it represents: the biology that emerges from our models is the combinatorial expression of all these automata doing their own little thing — just the way it is in the cell," says Gordon Webster, vice-president of biology at Plectix.

Complexity from simplicity

According to Edwards, the advantage of the Cellucidate approach is that a simple set of rules for each agent can result in complex biological behaviour when agents interact during the course of a simulation, unlike modelling in other formats, where the complexity has to be defined before a simulation can be executed. "The level of granularity also means that rules and agents can be easily recycled from one model to another," he notes. Like the GeneGo platform and the BioGRID initiative, Plectix relies on literature mining from various sets of experimental data to create the rules for a model system (see 'Playing by the rules', page 417).

"Mapping all interactions is important, but so is understanding the dynamics behind those interactions," says Snyder. To understand the dynamics of the information flow in cells, researchers not only need more knowledge of protein–protein interaction networks, but they also need to understand protein–DNA interactions, the effects of microRNAs and epigenetic changes on gene expression, and how other macromolecules such as metabolites affect the output of signalling networks. "It is the whole system together that determines the final output and activity," says Snyder.

Vidal thinks that technological improvements — especially in nanotechnology, to generate more data, and microscopy, to explore interaction inside cells, along with increased computer power — are required to push systems biology forward. "Combine all this and you can start to think that maybe some of the information flow can be captured," he says.

But when it comes to figuring out the best way to explore information flow in cells, Tyers jokes that it is like comparing different degrees of infinity. "The interesting point coming out of all these studies is how complex these systems are — the different feedback loops and how they cross-regulate each other and adapt to perturbations are only just becoming apparent," he says. "The simple pathway models are a gross oversimplification of what is actually happening."

Paul Nurse of Rockefeller University in New York wrote about understanding the cell's information flow last year[5]. He noted that "our past successes have led us to underestimate the complexity of living organisms", an oversight that is rapidly disappearing within the world of systems biology and will probably never happen again. ■

M. TYERS

Mike Tyers uses mass spectrometry to identify protein–protein interactions.

Nathan Blow is technology editor for *Nature* and *Nature Methods*.

1. Krogan, N. J. *et al. Nature* **440,** 637–643 (2006).
2. Gavin, A.-C. *et al. Nature* **440,** 631–636 (2006).
3. Tarassov, K. *et al. Science* **320,** 1465–1470 (2008).
4. Feret, J., Danos, V., Krivine, J., Harmer, R. & Fontana, W. *Proc. Natl Acad. Sci. USA* **106,** 6453–6458 (2009).
5. Nurse, P. *Nature* **454,** 424–426 (2008).
6. Vidal, M. *et al. Nature Methods* **6,** 39–46 (2009).
7. Vidal, M. *et al. Nature Methods* **6,** 47–54 (2009).
8. Vidal, M. *et al. Nature Methods* **6,** 83–90 (2009).
9. Vidal, M. *et al. Nature Methods* **6,** 91–97 (2009).