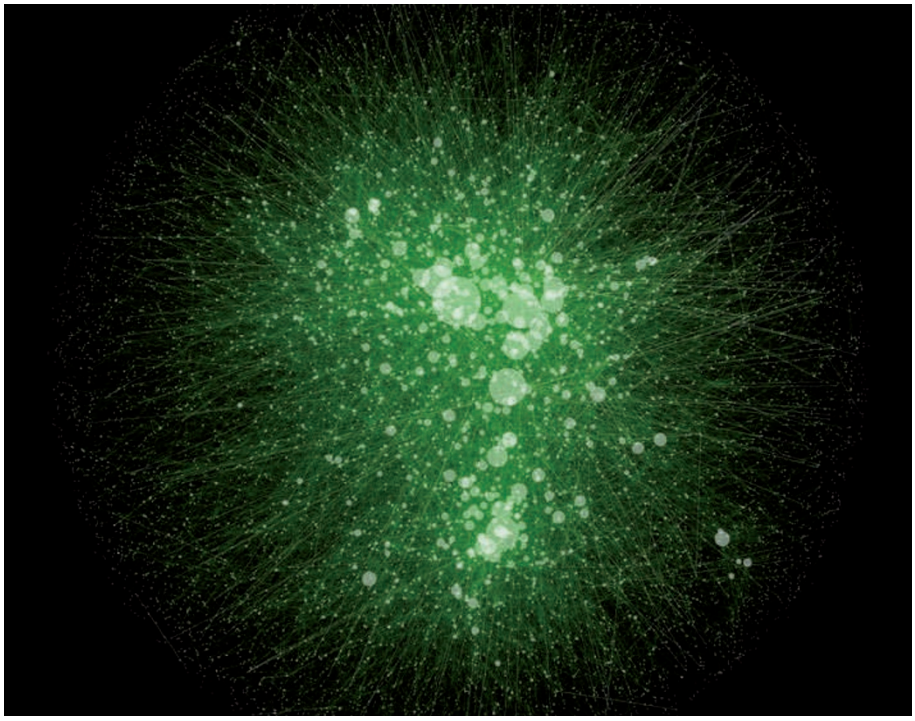


K. ONO/UC SAN DIEGO/CYTOSCAPE

The human interactome contains more than 100,000 protein interactions, only a fraction of which are known.

Interactome under construction

Developing techniques are helping researchers to build the protein interaction networks that underlie all cell functions.

BY LAURA BONETTA

An old adage says: "Show me your friends, and I'll know who you are." In the same way, finding interaction partners for a protein can reveal its function. To that end, researchers are now building entire networks of protein–protein interactions. Unlike biological pathways, which represent a sequence of molecular interactions leading to a final result — for example, a signalling cascade — networks are interlinked. Represented as starbursts of protein 'nodes' linked by interaction 'edges' to form intricate constellations, they provide insight into the mechanisms of cell functions. Furthermore, placing proteins encoded by disease genes into these networks will let researchers determine the best candidates for assessing disease risk and targeting with therapies.

"This is the next step after the Human Genome Project," says Trey Ideker, a systems biologist at the University of California, San Diego, and principal investigator at the National Resource for Network Biology, which provides open-source software for network visualization. "That effort identified 30,000 genes, but that is not the end goal. How the genes work in pathways and how these pathways function in disease states and development is the end goal. To accomplish this we will need to systematically map gene and protein interactions."

Unlike the genome, the interactome — the set of protein-to-protein interactions that occurs in a cell — is dynamic. Many interactions are transient, and others occur only in certain cellular contexts or at particular times in development. The interactome may be tougher to solve than the genome, but the information, researchers say, is crucial for a complete understanding of biology.

THE RIGHT PARTNERS

At any time, a human cell may contain about 130,000 binary interactions between proteins[1]. So far, a mere 33,943 unique human protein–protein interactions are listed on BioGRID (http://thebiogrid.org), a database that stores interaction data. Clearly, there is work to do.

There are two main approaches for detecting interacting proteins: techniques that measure direct physical interactions between protein pairs — binary approaches — and those that measure interactions among groups of proteins that may not form physical contacts — co-complex methods (see 'Tools for the search').

The most frequently used binary method is the yeast two-hybrid (Y2H) system[2]. It has variations involving different reagents, and has been adapted to high-throughput screening. The strategy interrogates two proteins, called bait and prey, coupled to two halves of a transcription factor and expressed in yeast. If the proteins make contact, they reconstitute a transcription factor that activates a reporter gene.

Another method for identifying binary interactions is luminescence-based mammalian interactome mapping (LUMIER), a high-throughput approach developed by Jeff Wrana at the Samuel Lunefeld Research Institute in Toronto, Canada. This strategy fuses Renilla luciferaze (RL) enzyme, which catalyses light-emitting reactions, to a bait protein, which is expressed in a mammalian cell along with candidate protein partners tagged with a polypeptide called Flag. Researchers use a Flag antibody to immunoprecipitate all proteins with the Flag tag, along with any that interact with them. Interactions between the RL-fused bait and the Flag-tagged prey are detected when light is emitted. Other binary methods include the mammalian protein–protein interaction trap and techniques based on proteome chips.

The most common co-complex method is co-immunoprecipitation (coIP) coupled with mass spectrometry (MS). In this approach, a protein bait is tagged with a molecular marker. Several types of tags are commercially available; each requires a distinct biochemical technique to recognize the tag and fish the bait protein out of the cell lysate, bringing with it any interacting proteins. These are then identified by MS.

In addition to these empirical methods, researchers have used computational techniques to predict interactions on the basis of factors such as amino-acid sequence and structural information. "People ask 'Why are you predicting interactions when you can just do the experiment?'" says Gary Bader, a bioinformatician at the University of Toronto. "But experimental techniques fail for some proteins."

FALSE READINGS

Every step of a procedure to detect protein–protein interactions — from the reagents used to the cell types and experimental conditions — influences the proteins that are identified. Two studies this year used similar methods to identify interacting proteins in transcription factors in embryonic stem cells[3,4]; there was incomplete overlap between the resulting data sets. "If you use the same protocol you will get reproducible lists of proteins. But different labs use different protocols, which affects the end result," says Raymond Poot, a cell biologist at Erasmus MC

hospital in Rotterdam, the Netherlands, and lead author of one of the studies.

In his protocol, Poot pulled interacting proteins from cells using nuclear extracts expressing different Flag-tagged transcription factors. He added a nuclease to his reactions to remove DNA and eliminate possible artefacts caused by proteins binding to it. "Transcription factors bind to DNA so you are likely to pull out DNA-binding factors that are not directly interacting," he explains. Purifying many different transcription factors with the same protocol also enabled the researchers to determine which interactions were most likely to be specific. For example, proteins that consistently co-purified with all transcription factors would be treated as unlikely to indicate a genuine interaction.

Calling out false positives — reported interactions that don't actually occur — and false negatives — interactions that do occur but are not picked up by the experimental protocol or are discarded — is one of the main challenges in the field. "Normally when you do a coIP followed by MS you will get hundreds of protein candidates interacting with any one bait," says Wade Harper, a cell biologist at Harvard Medical School in Boston, Massachusetts. "When you weed out all the stochastic and non-specific interactions you end up with many fewer proteins. Some proteins in large complexes might have 30–50 partners, others only 4–5."

One way in which researchers increase the accuracy of their results is to use more than one method (for example, Y2H plus LUMIER) to detect the interactions. But the definition of a 'real' interaction depends on the context. "Does a real interaction mean that two proteins interact if they are placed next to each other in a test tube, or that they must interact in a cell? Or does real mean that the interaction should have a biological function?" asks Ideker. Researchers can home in on functional interactions by combining data on interactions with other types of biological information, such as genetic interactions, protein localizations or gene expression. For instance, proteins whose genes are co-expressed are likely to interact with each other or to be part of the same complex or pathway.

Many tools are available on the web for integrating different types of information about a given protein or gene. One is GeneMANIA, developed by Bader's group in collaboration with Quaid Morris, a computational biologist also at the University of Toronto. A user enters the gene names into GeneMANIA; the program provides a list of genes that are functionally similar or have shared properties, such as similar expression or localization, and then displays a proposed interaction network, showing relationships among the genes and the type of data used to gather that information. The user can click on any node to obtain information about the gene and on any link to obtain information about their relationship (such as citations for any published studies or other sources of data). "It's like a Google for genetic and protein information," says Bader.

Other web-based interfaces that predict gene functions include STRING (http://string-db.org) developed at the European Molecular Biology Laboratory in Heidelberg, Germany. It hunts for protein interactions on the basis of genomic context, high-throughput experiments, co-expression and data from the literature.

KEEPING SCORE

To select real protein–protein interactions, Harper and some members of his lab, Matt Sowa and Eric Bennett, developed a software platform called CompPASS to assign confidence scores to an interaction detected by MS[5]. CompPASS takes data sets of interacting proteins (including those identified in experiments) and measures frequency, abundance and reproducibility of interactions to calculate the score.

This year, Harper used CompPASS to identify interactions among proteins involved in autophagy, the process by which cellular proteins and organelles are engulfed into vesicles and delivered to the lysosome to be degraded. Starting with 32 proteins known to have a role in autophagy, they identified 2,553 interacting proteins using coIP–MS. CompPASS then narrowed the list down to 409 high-confidence interacting proteins with 751 interactions[6].

Ideker's group used a different approach to map interactions among human mitogen-activated protein kinases (MAPKs), which respond to external stimuli and regulate cell function. Having used Y2H to identify more

Tools for the search

HYBRIGENICS

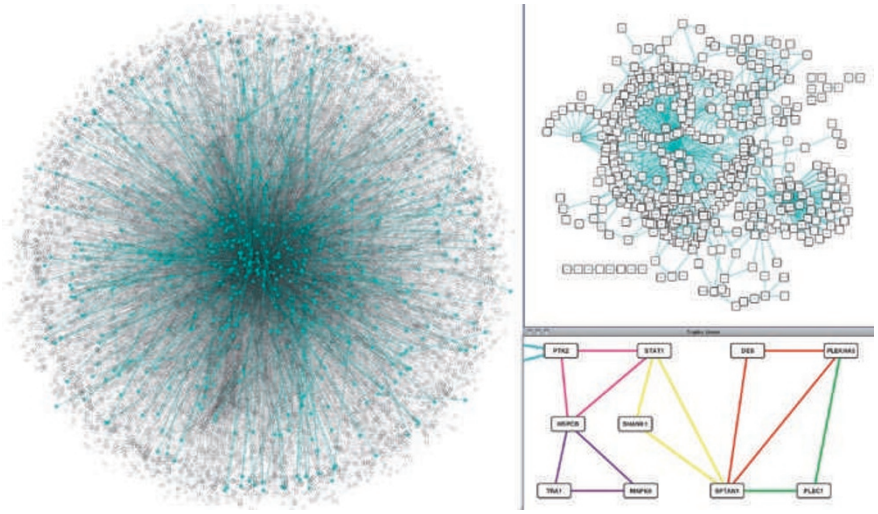
Methods such as the yeast two-hybrid system allow scientists to work out which proteins interact.

The two main methods for finding protein–protein interactions are the yeast two-hybrid (Y2H) system and co-immunoprecipitation followed by mass spectrometry. Several companies sell reagents for both approaches. Invitrogen of Carlsbad, California, sells the ProQuest Two-Hybrid System with Gateway Technology. This is based on Y2H, with modifications to decrease false-positive results and allow rapid characterization, says the company. Other firms provide vectors used to produce proteins with affinity tags, which can easily be immunoprecipitated along with other interacting proteins. A polypeptide tag called Flag is popular among researchers, and Sigma Aldrich of St Louis, Missouri, provides several Flag-genes for purchase. Promega in Madison, Wisconsin, has the HaloTag technology, in which a protein of interest is expressed in fusion with a tag protein engineered from a bacterial enzyme. This tag can be used to purify the protein, and any interacting with it, by binding to a resin. The tag is cleaved off using a protease.

For researchers who don't have the time or infrastructure to do the experiments, companies such as Hybrigenics in Paris and Dualsystems Biotech of Schlieren, Switzerland, offer Y2H-based screening. "We have complex libraries with ten times more independent clones than most other libraries, which we screen to saturation. And rather than screening full-length proteins, we screen for interactions with domains," says Etienne Formstecher, director of scientific projects and sales at Hybrigenics. "Full-length proteins can have some domains buried and not available to interact, at least in yeast where you may not have signals to unlock a closed protein conformation." A customer is given a list of proteins that interact with the protein of interest; it indicates which domains are making contact and provides a confidence score for each interaction.

Innoprot in Derio, Spain, provides an interaction service using tag-based purification designed for high-throughput analysis. And Invitrogen's ProtoArray Protein–Protein Interaction Service uses microarrays containing more than 9,000 human proteins to identify proteins that interact with any protein of interest. L.B.

K. ONO/UC SAN DIEGO/CYTOSCAPE



From a full network, researchers can zoom in on specific interactions that might be functionally relevant.

than 2,000 interactions among known MAPKs, Ideker used evidence including conservation of interactions among different species to winnow that down to a core network of 641 high-confidence interactions[7].

For some of the proteins there was no previous evidence of interactions with MAPKs. Ideker and his colleagues knocked down the expression of these proteins using RNA interference, then looked for the effect of the knockdowns on proteins known to be activated by MAPKs. This allowed them to confirm that about one-third of their interactions had a role in MAPK signalling.

These methods are helping to weed out false positives and provide associated confidence scores, but the problem of false negatives persists. "With these assays we try to get false positives down to zero. The hit you take is on false negatives. So now you can be highly confident of your data but you are probably probing only about 20% of the interactome," says Ideker. "We would like to get every interaction but we do not get even close with current technologies."

New methods may become available to identify interactions that escape detection by current techniques (see 'Real-time analysis'). In the meantime, one way to address the problem is to combine procedures for detecting interactions, each sampling a different portion of the interactome. The interaction data obtained in an experiment can also be combined with that available in public databases, thus providing a more complete picture, says Bader.

FROM DATA TO NETWORKS

Protein–protein interactions are only the raw material for networks. To build a network, researchers typically combine interaction data sets with other sources of data. Primary databases that contain protein–protein interactions include DIP (http://dip.doe-mbi.ucla.edu), BioGRID, IntAct (www.ebi.ac.uk/intact) and MINT (http://mint.bio.uniroma2.it). These databases have committed to making records available through a common language called PSICQUIC, to maximize access.

Other types of data that can be combined with protein–protein interactions include information on gene expression, cellular co-localization of proteins (based on microscopy), genetic information, metabolic and signalling pathways, and data from high-throughput assays.

"One challenge computationally is integrating heterogeneous data sets to build a network model," says Ilya Shmulevich, a professor at the Institute for Systems Biology in Seattle, Washington. The second challenge is to decide on a modelling approach. "It will depend on what kind of data you have available and how you will be using the model," says Shmulevich.

Several bioinformatic tools have been developed to model and represent networks. The most widely used ones are associated with Cytoscape (www.cytoscape.org), an open-source program for visualizing networks and for integrating them networks with other types of data. Several Cytoscape plug-ins allow users to download and explore databases.

Commercial packages with similar functions include MetaCore from GeneGO in St Joseph, Michigan; Pathway Analysis from Ingenuity Systems in Redwood City, California; and Pathway Studio from Ariadne Genomics in Rockville, Maryland. These can access public sources of data as well as the company's proprietary databases. "One of the unique features of Pathway Studio is the openness of our system and the ability to integrate many different kinds of data," says David Denny, director of marketing and product management at Ariadne.

GUILT BY ASSOCIATION

One reason for developing networks is to help assign functions to proteins through guilt by association. But "a huge slice of the proteome consists of proteins that no one knows what they do or interact with", says Benjamin Cravatt, a chemical physiologist at the Scripps Research Institute in San Diego, California.

For proteins not yet assigned to a portion of the human interaction network, Cravatt's group developed a technology for assigning protein functions by exploiting an interaction between enzymes and chemical reagents dubbed activity-based probes. These probes consist of a reactive group that binds the active sites of many members of an enzyme family, and a reporter tag that is used for the detection and identification of the probe-labelled enzymes[8].

Because these probes bind only to enzymes that are active, they can give insights into the enzymes' functions. For example, if a probe binds to a set of enzymes in a cancer cell but not in a normal cell, it means that these enzymes become more active in the cancer cell and so may have a role in cell growth. The activity probes can also serve as assays for the discovery of inhibitors for a particular enzyme, which may help researchers to understand the role of that enzyme. "You can develop an inhibitor for an enzyme before ever knowing what the actual substrate is," says Cravatt.

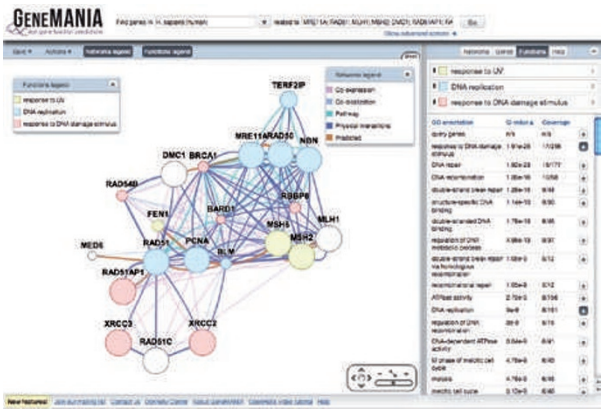
This year, he developed another strategy that not only determines differences in enzyme activities in different cells, but also pinpoints where in the protein these differences occur, providing a more quantitative measure of the differences[9]. The activities of many families of enzymes are regulated or fine-tuned by cysteine modifications. By looking specifically for changes in cysteine modifications across the proteome, he found 'hyper-reactive' cysteine residues in several proteins of unknown function, which suggests that they probably have roles in signalling pathways.

"One challenge is integrating heterogeneous data sets."
Ilya Shmulevich

One challenge in defining protein–protein interaction networks is that interactions vary depending on the type of cell and the cellular environment. For example, Wrana mapped the protein–protein interaction network for TGF-β, a growth factor that regulates cell functions, and found that two proteins that pass on the signals from the factor inside the cell — Smad2 and Smad4 — interact with one another only when the cells are stimulated with TGF-β. If the cells are not stimulated, these two proteins don't come into contact[10].

Bennett, Harper and Steven Gygi, a cell biologist also at Harvard Medical School, developed a proteomics platform centred around a technology called multiplex absolute quantification (AQUA) to look at dynamic changes in protein interaction networks. AQUA uses synthetic peptides that contain stable isotopes as internal standards for the native peptides that are produced when proteins from a cell

lysate are digested. Using tandem MS, researchers can compare the levels of native and synthetic peptides in a cell to obtain a measure of the amount of native proteins present. Synthetic peptides can also be prepared with modifications, such as extra phosphate groups, to measure the number of post-translationally modified proteins. "We are pursuing the dynamics of protein networks by quantifying changes in the amount of proteins present in specific protein complexes," says Harper. "Techniques such as AQUA provide an accurate and sensitive measure of how the stoichiometry of components within complexes that make up a network are altered in response to a stimulus."



Web tools such as GeneMANIA integrate data on a protein or gene.

The team used the approach to describe the rearrangements that occur in the protein network of cullin-RING ubiquitin ligases, enzymes that regulate protein turnover, under various cellular conditions[11].

DEVELOPMENTS IN DIAGNOSTICS

Changes in protein–protein interaction networks may provide information about the mechanisms of disease. Last year, Wrana applied the network approach to the diagnosis of breast cancer. He used microarrays to measure genome-wide protein expression in the tumours of people with breast cancer, and then overlaid the expression data on the network diagram of the human interactome.

Wrana had noted that 'hub' proteins, defined as those that interact with more than four others, can be grouped into two categories depending on whether they are expressed at the same time as the proteins with which they interact. When they looked at breast-cancer samples, Wrana and his colleagues found that certain hub proteins were in a different category in breast-cancer patients with a good prognosis than in those with a poor prognosis.

Thus, by overlaying the expression pattern of a cancer cell from an individual patient onto the human interactome network, Wrana could predict a patient's prognosis. "We found that the detection of global changes in network organization is more predictive of outcome than is gene expression alone," says Wrana. "We have now applied this method to other tumour models and obtained similar results."

KAYAK (kinase activity assay for protein profiling) is another approach to developing diagnostic tools for cancer on the basis of the functional consequences of the interaction between a protein, in this case a kinase, and its substrate. In this method, up to 90 peptide substrates for kinases are used to simultaneously measure the addition of phosphate groups to proteins in a cell lysate — in essence providing a 'phosphorylation signature' for that particular cell. "The readout is so sensitive and so quantitative that even small differences are teased out," says Gygi, who helped to develop the method[12].

According to Gygi, the biggest application of KAYAK might be in tumour classification. "Biopsies or excised tissues can be profiled for kinase activities with pinpoint accuracy. These patterns could contribute towards personalized drug treatments based on dysregulated kinase pathways," he says.

The combination of different types of data and technologies should continue to fill in the empty spaces of the current human interactome map. The picture may never be complete, but it will continue to provide insights into cellular mechanisms of health and disease. "I think that the network we have is dense enough for us to start doing studies to classify disease states," says Wrana. "As the networks become better and coverage improves, the accuracy of diagnosis will also improve." ■

Laura Bonetta *is a freelance science writer based in Garrett Park, Maryland.*

1. Venkatesan, K. *et al. Nature Meth.* **6,** 83–90 (2009).
2. Fields, S. & Song, O.-K. *Nature* **340,** 245–246 (1989).
3. van den Berg, D. L. C. *et al. Cell Stem Cell* **6,** 369–381 (2010).
4. Pardo, M. *et al. Cell Stem Cell* **6,** 382–395 (2010).
5. Sowa, M. E., Bennett, E. J., Gygi, S. P. & Harper, J. W. *Cell* **138,** 389–403 (2009).
6. Behrends, C., Sowa, M. E., Gygi, S. P. & Harper, J. W. *Nature* **466,** 68–76 (2010).
7. Bandyopadhyay, S. *et al. Nature Meth.* **7,** 801–805 (2010).
8. Nomura, D. K., Dix, M. M. & Cravatt, B. F. *Nature Rev. Cancer* **10,** 630–638 (2010).
9. Weerapana, E. *et al. Nature* **468,** 790–795 (2010).
10. Barrios-Rodiles, M. *et al. Science* **307,** 1621–1625 (2005).
11. Bennett, E. J. *et al. Cell* (in the press).
12. Yonghao, Y. *et al. Proc. Natl Acad. Sci. USA* **106,** 11606–11611 (2009).
13. Uemura, S. *et al. Nature* **464,** 1012–1017 (2010).

Real-time analysis

In November, Pacific Biosciences of Menlo Park, California, commercially released its third-generation DNA-sequencing platform, based on its single-molecule, real-time (SMRT) technology. A single DNA polymerase bound to a DNA template is attached to a tiny chamber illuminated by lasers, and nucleotides labelled with coloured fluorophores are introduced to it. As the polymerase incorporates them, each base is held for a few microseconds, while the fluorophore emits coloured light corresponding to the base identity. SMRT technology could also be used to analyse biomolecules other than DNA, and could become a common tool for detecting protein interactions, with some unique features. "This technology can detect relatively weak interactions," says Jonas Korlach, a scientific fellow at Pacific Biosciences, adding that it could pick out interactions that happen so quickly that they can't be identified by current methods.

As a step towards such applications, Joseph Puglisi, a structural biologist at Stanford University School of Medicine in California, and his group, with scientists at Pacific Biosciences, observed transfer RNAs binding to single ribosomes in real time[13]. In an unpublished follow-up, Puglisi's group has used SMRT technology to watch interactions between transfer RNAs, ribosomes and protein factors to determine how the translation machinery synthesizes proteins. "We have just seen the tip of the iceberg in terms of applications," says Korlach. L.B.

PACIFIC BIOSCIENCES

Future SMRT systems could reveal interactions.