Searching high and low for interactions

High-throughput protein-protein interaction studies typically are met with skepticism from scientists engaged in hypothesis-driven research. But is it fair?

If you were given a choice between a large protein-protein interaction data set obtained by high-throughput experiments and another data set extracted from a collection of small-scale experiments via literature mining, which one would you trust more?

An anecdotal poll indicates that many scientists tend to attribute some intrinsic qualities to low-throughput experiments–qualities that they find sorely missing in the high-throughput ones. Despite the milestones achieved over the past few years, high-throughput interactome studies continually encounter a good deal of skepticism. But a lot of progress has been made since the first rough interactome drafts. Although considering them with a grain of salt is still appropriate, it is unreasonable at this point to simply dismiss high-throughput studies as unreliable.

Certainly small-scale studies of individual complexes offer options for follow-up, confirmation and validation that largely surpasses the present possibilities afforded by high-throughput protein-interaction screens. And yet, not all small-scale studies are created equal, and validation is not automatically associated with small-scale studies, just as it is not automatically excluded from large-scale efforts. On the contrary, validation is becoming an increasingly important part of large-scale studies.

For example, consider the yeast two-hybrid assay, often criticized for its high propensity to yield both experimental and biological false positive results. It has become the norm for interactome studies using this assay to include sophisticated quality control schemes involving validation by different techniques, along with a thorough estimate of technical false positive and negative rates. Biological false positives remain a caveat more difficult to pin down, but yeast two-hybrid screens, if well controlled, offer a unique opportunity to discover all putative pairwise interactions within a group of proteins, providing a strong scaffold upon which other approaches may build.

As more comprehensive data sets become available, investigators can begin to validate high-throughput results by replication. For example, the two most comprehensive yeast interactome studies to date were published last year, both carried out by systematic affinity purification of protein complexes. These two data sets have now been merged and reanalyzed with a specially designed scoring system that takes into consideration, for each interaction, the collective amount of evidence—and counter-evidence—that could be gathered from the two independent studies. Most interestingly, the resulting consolidated data set compares favorably to similar data extracted from a repository of interactions identified in small-scale experiments (see Research Highlights p389).

This should not come as a surprise. High-throughput experiments also have advantages. In particular, their measurements are performed in a standardized way, and this notably allows the systematic collection of both positive and negative results. In this manner, high-throughput studies avoid the biological bias—or 'inspection' bias—that can affect some hypothesis-driven studies of individual complexes.

Standardized data sets are also more amenable to analysis than disparate information curated from the literature. Bioinformaticians have made progress in developing strategies that take into account the limitations of underlying biological assays while capitalizing on the comparability of large-scale data sets and the multiple lines of positive and negative evidence they contain. Statisticians are also devising ways of integrating not only independent experimental replication but also different types of data. Physical interaction data obtained by yeast two-hybrid and affinity purification will increasingly be combined with colocalization data, coexpression data from gene expression profiling, genetic interaction data and Gene Ontology annotations. This leads to a probabilistic view of interactomes: instead of being defined as 'interacting' or 'noninteracting', each protein pair is placed in a spectrum of confidence levels. As more data become available, the confidence score of each particular interaction may change, thus refining the quality of the interactome description.

Advocates of high-throughput interactomics approaches have compared the present situation to the development of high-throughput sequencing, pointing out that in the early days of genome sequencing the quality achieved for individual genes was not as good as what could be obtained with slab gels. It is only as the technology developed, as researchers accumulated multiple sequence coverage and as they developed the appropriate data-analysis algorithms, that the quality of genome sequencing increased.

Now seems an appropriate time to abandon the idea that the body of information on protein interactions available in the literature is superior to results of high-throughput experiments. The two are becoming increasingly comparable; both containing interactions validated to varying confidence levels. High-throughput data sets, however, because of their systematic and standardized nature, can be improved further, and they have the potential to reveal aspects of biology that are not accessible to conventional hypothesis-driven research.