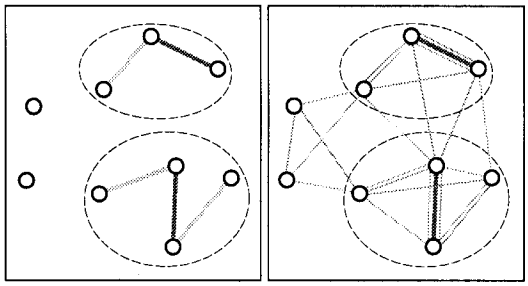
Integrating Interactomes

Mark Gerstein, Ning Lan, Ronald Jansen

With the human genome sequence as an intellectual inspiration and practical scaffold, scientists are ready to perform experiments on all genes. Integrating the resulting genomewide information into useful definitions of protein function is a huge challenge. Exactly what form such functional definitions will take is still debatable, but comprehensive networks of protein-protein interactions, or interactomes, should prove valuable in helping to shape them.

Enhanced online at www.sciencemag.org/cgi/content/full/295/5553/284

On page 321 of this issue, Tong *et al.* (*1, 2*) describe a systematic approach for identifying protein-protein interaction networks in which different peptide recognition domains participate. They break new ground in the way they combine "orthogonal" (that is, fundamentally different) sets of genomic information. Specifically, they study the intersection of two different interactomes. The first is derived from screening phage-display peptide libraries to find consensus sequences in yeast proteins that bind to particular peptide recognition domains. The resulting network connects proteins with recognition domains to those containing the consensus. This network partially defines binding sites in some of the proteins and represents a clever use of phage display technology. The second network is derived from experimentally testing each peptide recognition module, using the yeast two-hybrid technique, for association with possible protein-binding partners. Tong *et al.* apply their approach to determine interacting partners for SH3 domains in yeast proteins. These domains make good targets because of their prevalence and involvement in a number of important biological processes, from cytoskeleton reorganization to signal transduction.



Overlapping nets. Two different extremes in integrating interactomes. The combined network on the left is the union of those interactomes with low false-positive but high false-negative rates, whereas the combined network on the right is the intersection of interactomes with low false-negative but high false-positive rates. Circles represent proteins; links, interactions; and dotted lines, known associations. Thicker links indicate lower false-positive rates. More effective rules for combining networks than union and intersection take into account the different error rates associated with each link type.

The power of Tong *et al.*'s strategy (particularly for reducing noise) becomes manifest when interpreting large genomic data sets. One fallacy in dealing with genomic data sets is ascribing too much meaning to individual data points. Many data sets (for example, gene expression profiles) contain so much noise that it can be difficult to draw reliable conclusions for specific genes. These data sets still offer much useful information statistically, in terms of broad trends, but they are useful only insofar as the data can be aggregated. This can be simply achieved by combining replicates of an experiment, but such a process does not remove systematic errors. It is also possible to collect many individual measurements on different proteins into aggregate "proteomic classes," for example, functional categories, and to compare these (*3, 4*).

The new work points to perhaps the most powerful approach: interrelating and integrating orthogonal information. In the abstract, it is easy to demonstrate that combining independent data sets results in a lower error rate overall. For instance, combining three independent binary-type data sets with error rates of 10% (for false positives and negatives) reduces the overall error rate to 2.8% (for positives and negatives) (*5–7*). Moreover, interrelating two different types of whole-genome data also enables one to discover potentially important but not obvious relationships—for example, between gene expression and the position of genes on chromosomes, or between gene expression and the subcellular localization of proteins (*8, 9*).

There have been a number of previous attempts to interrelate information from different genomic data sets. For instance, gene expression profiles were initially analyzed by a variety of supervised and unsupervised methods—hierarchical trees, k-means, self-organizing maps, and support-vector machines—and compared with protein func-

CREDIT: (TOP) ADAPTED FROM (8)

The authors are in the Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA. E-mail: mark.gerstein@yale.edu

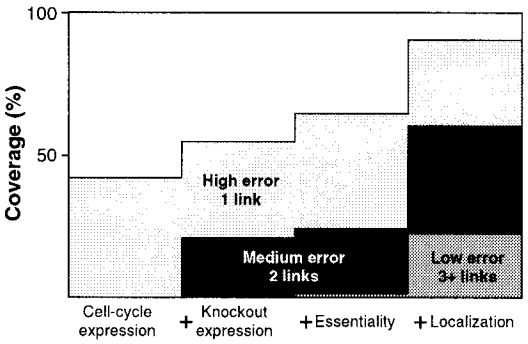
tion categories (*10–14*). Gene expression data were also compared with data sets describing transcription factor binding sites, protein families, protein-protein interactions, and protein abundance (*3–6*, *15–19*). In a shorthand sense, much of this can be thought of as interrelating the transcriptome (population of mRNA transcripts) with other "omes" such as the proteome, translatome, secretome, and interactome (*3*).

There are considerably fewer examples of the synthesis of more than two types of genomic information. One initial attempt combined gene expression correlations, phylogenetic profiles, and patterns of domain fusion to predict protein function (*20–22*). Bayesian statistics were used to integrate gene expression, "essentiality" (the degree to which a gene is essential for survival), and sequence motif data into a uniform framework for the prediction of protein subcellular localization (*20*). Tong *et al.*'s strategy of overlapping interactomes presents a new type of synthesis. It is particularly effective in that their two data sets are orthogonal in many respects. Phage display is based on in vitro binding of short peptides, whereas the two-hybrid approach assays in vivo binding between full-length proteins. Moreover, the phage display network is computationally predicted but uses relatively unambiguous consensus sequences, whereas the two-hybrid network is experimentally derived but suffers from appreciable false positives (*23*, *24*).

From a data-mining standpoint, the heterogeneous character and variable quality of whole-genome information makes integration tricky. Consider combining "orthogonal" interactome data sets, such as attempted by Tong *et al.*, in a general sense. How might one proceed formally? There are two extremes (see figure, previous page). At one extreme, the data sets have low false-negative but high false-positive error rates. That is, each experiment almost never misses real interactions but also finds many spurious ones. In this situation, the benefit of integration comes from intersection: Only interactions common to all are accepted, thus lowering the combined error rate. Tong *et al.*'s approach fits this to some degree. At the other extreme are data sets with few false positives but low coverage of the space of interactions. The benefit of integration then comes from the union: Any interaction found in at least one data set is accepted. An earlier interactome analysis followed this to some degree (*25*).

In most practical situations, the optimal way to integrate data sets is somewhere between these extremes. The task is to combine data sets with varying error rates and coverage. Accordingly, the rules for identifying positives become more complicated. Instead of simple unions or intersections, different combinations of positive and negative signals from the data sets should be considered, taking into account their relative false-positive and -negative rates.

A practical illustration of the power of interrelating genomic data for yeast (see figure, this page) shows the degree to which one can find protein-protein associations in known protein complexes (*5*, *6*, *26*) by stepwise integration of increasing amounts of orthogonal genomic information. We start by considering associations that can be found from gene expression correlations over the cell cycle (*27*); then we incorporate those derived from a second but different microarray experiment, which provides a series of gene expression changes after specific genes have been knocked out (*28*). Finally, we add associations predicted from genomic measurements of essentiality and localization (*20*, *26*, *29*, *30*). As we integrate more information, the total number of correctly identified interactions rises (especially for the union of the predicted associations). Simultaneously, the error rate decreases. Moreover, if we focus just on the intersection of the predicted associations, the error rate falls even more.



A net profit from intergation. Integrating progressively more orthogonal information identifies more and more associations (*5–7*). From the known complexes in yeast, there are 8250 protein-protein associations (*26*). The *y* axis shows the percentage of these identified by disparate genomic data (that is, coverage). The *x* axis shows the progressive addition of genomic data. The first two bars represent the protein associations with the most significant expression correlation in two different microarray sets (*27, 28*). The next two represent adding the associations predicted because both proteins were similarly essential for cell survival ("essentiality") or had similar subcellular localization (*20, 29, 30*). The color shading on the bars roughly indicates false-positive rates throughout the integration. Although it is reasonable that associated components of complexes will have correlated expression and similar localization and "essentiality," this is only weakly predictive, generating many spurious positives. Consequently, the "weak links" case in the right hand panel of the top figure applies, and the shading indicates how intersection lowers the error rate.

A major challenge will be devising uniform frameworks for integrating information from both high-throughput and traditional biochemical approaches. One aspect of this will be developing better databases for storing and querying heterogeneous information. In particular, databases will need to be more precise in their treatment of errors and also interface better with the information in journals. Another aspect will be to develop data-mining strategies that can operate with these databases, integrating many different genomic features into results pertinent to biology. Genomic features can be of very different character (from hundreds of "Booleans" for interactions, to tens of thousands of real-number vectors for expression profiles), and a central issue in integration is determining how to weight each feature relative to the others. In this regard, some machine-learning techniques, such as Bayesian networks and decision trees, are quite powerful, whereas others (for example, support-vector machines) are more problematic.

Finally, we will need to come up with a more systematic definition of gene function, the ultimate aim of proteomic investigation. To many scientists, what constitutes "function" is a phrase or name, often in nonsystematic terminology, "ATPase" or "suppressor of white apricot" for example. Such descriptions are sufficient for single-molecule work but cannot be scaled up to the genomic level. More systematic attempts have been made to place proteins within a hierarchy of standard functional categories or to connect them in overlapping networks of varying types of association (*26*, *31*, *32*). These networks can obviously include protein-protein interactions, the subject of Tong *et al.*'s work. More broadly, they can include pathways, regulatory systems, and signaling cascades. How far are we able to go with this network approach? Perhaps in the future the systematic combination of networks may provide for a truly rigorous definition of protein function.

References

1. A. Tong *et al.*, *Science* **295**, 321 (2002).
2. Interaction data from Biomolecular Interaction Network Database (www.binddb.org).
3. D. Greenbaum *et al.*, *Genome Res.* **11**, 1463 (2001).
4. R. Jansen, M. Gerstein, *Nucleic Acids Res.* **28**, 1481 (2000).
5. J. Qian *et al.*, *J. Mol. Biol.* **314**, 1053 (2001).
6. R. Jansen *et al.*, *Genome Res.* **12**, 37 (2002).
7. Details at http://genecensus.org/integrate/interactions.
8. B. A. Cohen *et al.*, *Nature Genet.* **26**, 183 (2000).