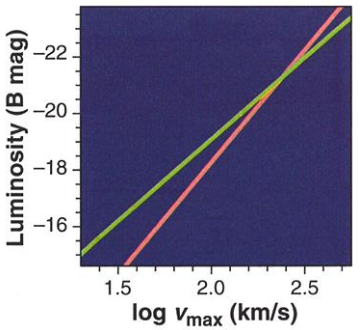
nosity during the same period. Explaining this result constitutes a challenge for different models of galaxy evolution (*8*).

The exploration of distant galaxies requires accurate and well-defined projects. In the past, many surveys were concerned with mapping the whole sky. In contrast, the surveys of the future will have to concentrate on well-defined areas at maximum resolution and with a range of instruments. In this spirit, the Great Observatories Origins Deep Survey (GOODS) aims to survey a small area of the sky with several major astronomical facilities (including the Chandra X-ray Telescope, the Hubble Space Telescope, and the ESO VLT telescope), covering the entire range of wavelengths at our disposal (*9*). The total area to be surveyed is only 300 square arc min—similar to that subtended by the full Moon—but large enough to give us an idea of what happened at the beginning of the universe.



Galaxies near and far. The slope of the Tully-Fisher relation for 1200 local spiral galaxies (pink) is steeper than that for 60 spiral or irregular galaxies at intermediate redshift (green). The result provides insights into how galaxies of different mass and luminosity may have evolved over time.

References

1. The Mass of Galaxies at Low and High Redshift, workshop organized by ESO and the Universitäts-Sternwarte München, Venice, 24 to 26 October 2001; see www.eso.org/gen-fax/meetings/gmass2001.
2. M. I. Wilkinson, N. W. Evans, *Mon. Not. R. Astron. Soc.* **310**, 645 (1999). GAIA stands for Global Astrometric Interferometer for Astrophysics.
3. R. P. Olling, M. Merrifield, *Mon. Not. R. Astron. Soc.* **311**, 361 (2000).
4. _____, *Mon. Not. R. Astron. Soc.* **326**, 164 (2001).
5. S. S. McGaugh, V. C. Rubin, W. J. G. de Blok, *Astron. J.* **122**, 2381 (2001).
6. W. J. G. de Blok, S. S. McGaugh, V. C. Rubin, *Astron. J.* **122**, 2396 (2001).
7. D. Merritt, L. Ferrarese, in *The Central Kpc Starbursts and AGN*, J. H. Knapen *et al.*, Eds. [Astronomical Society of the Pacific (ASP) Conference Series, ASP, San Francisco, in press]; see http://xxx.lanl.gov/abs/astro-ph/0107134.
8. B. L. Ziegler *et al.*, in preparation; see http://xxx.lanl.gov/abs/astro-ph/0111146.
9. R. Fosbury *et al.*, *Messenger* **105**, 40 (2001).

PERSPECTIVES: PROTEOMICS

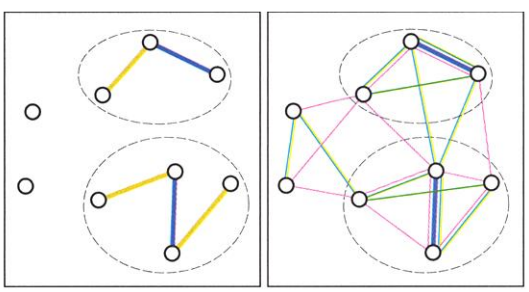
Integrating Interactomes

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 (*3–6*).

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% reduces the overall error rate to 2.8% (for both false positives and negatives) (*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

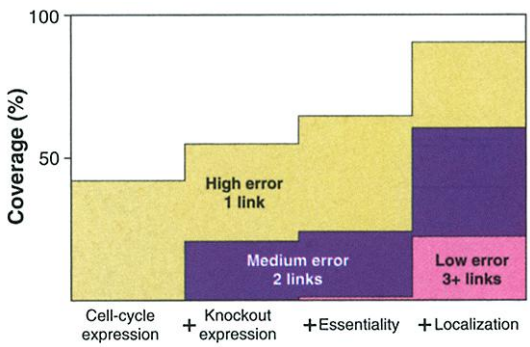
tion categories (*10–14*). Gene expression data were also compared with data sets describing transcription factor binding sites, protein families, protein-protein interactions, and protein abundance (*3–6*, *15–20*). In a shorthand sense, much of this can be thought of as interrelating the transcriptome (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 (*21*, *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 the 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 the 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 integration. 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 previous figure mostly applies, and the shading indicates how intersection lowers the error rate.

A future challenge will be to devise uniform frameworks for integrating information from both high-throughput and traditional biochemical approaches. One aspect of this will be to develop 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, such as "ATPase" or "suppressor of white apricot." 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. H. Y. Tong *et al.*, *Science* **295**, 321 (2002); published online 13 December 2001 (10.1126/science.1064987).
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).

9. A. Drawid *et al.*, *Trends Genet.* **16**, 426 (2000).
10. P. Tamayo *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 2907 (1999).
11. M. B. Eisen *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14863 (1998).
12. S. Tavazoie *et al.*, *Nature Genet.* **22**, 281 (1999).
13. M. Gerstein, R. Jansen, *Curr. Opin. Struct. Biol.* **10**, 574 (2000).
14. M. Brown *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 262 (2000).
15. H. Ge *et al.*, *Nature Genet.* **29**, 482 (2001).
16. F. Roth *et al.*, *Nature Biotechnol.* **16**, 939 (1998).
17. S. Gygi *et al.*, *Mol. Cell. Biol.* **19**, 1720 (1999).
18. B. Futcher *et al.*, *Mol. Cell. Biol.* **19**, 7357 (1999).
19. A. Brazma *et al.*, *Genome Res.* **8**, 1202 (1998).
20. A. Drawid, M. Gerstein, *J. Mol. Biol.* **301**, 1059 (2000).
21. E. Marcotte *et al.*, *Science* **285**, 751 (1999).
22. E. Marcotte *et al.*, *Nature* **402**, 83 (1999).
23. T. Ito *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 4569 (2001).
24. P. Uetz *et al.*, *Nature* **403**, 623 (2000).
25. B. Schwikowski *et al.*, *Nature Biotechnol.* **18**, 1257 (2000).
26. H. W. Mewes *et al.*, *Nucleic Acids Res.* **28**, 37 (2000).
27. R. J. Cho *et al.*, *Mol. Cell* **2**, 65 (1998).
28. T. R. Hughes *et al.*, *Cell* **102**, 109 (2000).
29. E. A. Winzeler *et al.*, *Science* **285**, 901 (1999).
30. P. Ross-Macdonald *et al.*, *Nature* **402**, 413 (1999).
31. D. Eisenberg *et al.*, *Nature* **405**, 823 (2000).
32. M. Ashburner *et al.*, *Nature Genet.* **25**, 25 (2000).

Do X Chromosomes Set Boundaries?

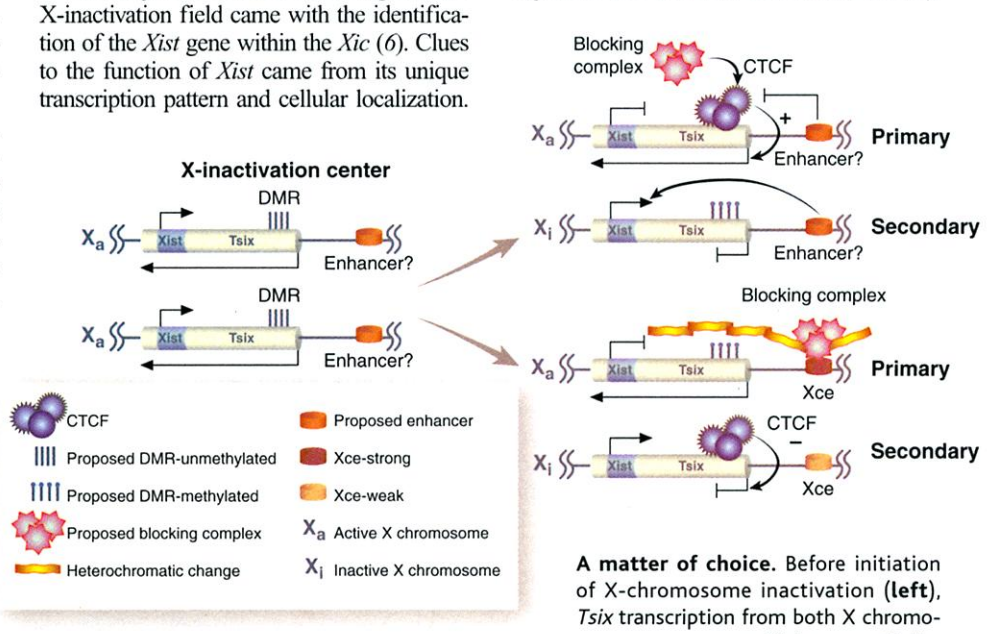
Ivona Percec and Marisa S. Bartolomei

Sexually dimorphic organisms employ the services of epigenetics—heritable changes in gene expression that are independent of DNA sequence—to balance genetic differences between the two sexes. A superb model of this relationship, X-chromosome inactivation, has evolved uniquely in mammals to ensure equal gene dosage between females, who have two X chromosomes, and males, who have only one X. This precise pathway results in the silencing of the majority of genes on one X chromosome early in female development. This outcome requires a female cell to undergo a highly orchestrated set of events when it differentiates. A cell must count the X chromosomes, choose one X to inactivate (usually in a random manner), initiate and propagate chromosome-wide silencing, and finally maintain this inactive state throughout subsequent cell divisions (*1*). Shortly after the discovery of X inactivation by Mary Lyon in 1961, geneticists hypothesized that cis-acting factors (acting on the same chromosome) encoded by the X must be important in this process. Likewise, trans-acting factors (acting on different chromosomes) encoded by chromosomes other than the X or Y were presumed to be equally important (*2*). Yet until recently, all known regulators of X inactivation were cis-acting elements residing on the X chromosome. The drought surrounding the identification of trans-acting factors has now ended. According to Chao *et al.* (*3*) on page 345 of this issue, the insulator and transcription regulator CTCF is a key trans-acting factor in the X-inactivation pathway.

Early studies on X inactivation demonstrated that a region of the X chromosome, designated the X-inactivation center (*Xic*), is required for silencing of adjacent sequences (*4*). As a result, a chromosomal fragment containing the *Xic* can become inactive, whereas one that does not, by default, must remain active. In addition to delineating the *Xic* as the principal cis-acting silencing center, early experiments uncovered a genetic element within the *Xic* that affects X-chromosome choice in the mouse (*5*). Alleles of this element, named the X controlling element (*Xce*), vary in strength such that a strong *Xce* allele is more likely to reside on an active X chromosome than a weak *Xce* allele. Surprisingly, *Xce* has escaped molecular identification.

The major molecular breakthrough for the X-inactivation field came with the identification of the *Xist* gene within the *Xic* (*6*). Clues to the function of *Xist* came from its unique transcription pattern and cellular localization. *Xist*, a gene that does not encode a protein, is transcribed from the inactive X chromosome (X_i) and is silent on the active X chromosome (X_a). It codes for a large untranslated RNA that coats the X_i. Genetic experiments have demonstrated that *Xist* is required for initiation and promulgation of silencing, and that it is involved in X-chromosome choice (*1*). These findings invoked a compelling molecular model of initiation and propagation events, with the *Xist* RNA acting as the major inactivating element. Despite this progress, molecular candidates directing the initial events of counting and selection remained elusive.

Studies of the antisense gene *Tsix*, the most recent addition to the cis-acting family of factors within the *Xic*, have begun to illuminate these early events (*7*). *Tsix* overlaps with *Xist*, but is transcribed from the antisense strand. Like *Xist*, *Tsix* codes for an untranslated RNA, yet contrary to *Xist*, *Tsix* is transcribed from the X_a. This pattern suggests that the two genes are coordinately regulated and that *Tsix* blocks *Xist* activity.



A matter of choice. Before initiation of X-chromosome inactivation (**left**), *Tsix* transcription from both X chromosomes suppresses *Xist* gene activity, preventing X-chromosome silencing. During X-chromosome choice, CTCF may bind to the future X_a as a primary event preventing *Xist* transcription (**top right**). In this scenario, suppression of *Xist* by CTCF could be achieved by direct activation of its repressor, *Tsix*, or by blocking access to putative enhancers located downstream. Alternatively, a blocking complex may bind to the future X_a as a primary event inducing heterochromatic changes within the *Xic*, including methylation and suppression of *Xist* (**bottom right**). In this scenario, CTCF binds to the future X_i as a secondary event and either directly represses *Tsix*, or blocks *Tsix*'s access to enhancers close to *Xist*. The enhancers have not yet been identified, and their location is speculative.

The authors are at the Howard Hughes Medical Institute and Department of Cell and Developmental Biology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA. E-mail: bartolom@mail.med.upenn.edu, ipercec@mail.med.upenn.edu