Acknowledgements
We thank the Deutsche Forschungsgemeinschaft for financial support. S.H. was supported by a grant of the Landesgraduiertenförderung Baden-Württemberg.

References

1 H.K. Moghadam *et al.* Organization of Hox clusters in rainbow trout (*Oncorhynchus mykiss*): a tetraploid model species. *J. Mol. Evol.* (in press)
2 Aparicio, S. *et al*. (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* 297, 1301–1310
3 Jaillon, O. *et al*. (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431, 946–957
4 Amores, A. *et al*. (2004) Developmental roles of pufferfish Hox clusters and genome evolution in ray-fin fish. *Genome Res.* 14, 1–10
5 Chiu, C-h. *et al*. (2004) Bichir *Hox*A cluster sequence reveals surprising trends in ray-finned fish genomic evolution. *Genome Res.* 14, 11–17
6 Powers, T.P. and Amemiya, C.T. (2004) Evolutionary plasticity of vertebrate Hox genes. *Curr. Genomics* 5, 459–472
7 Powers, T.P. and Amemiya, C.T. (2004) Evidence for a *Hox*14 paralog group in vertebrates. *Curr. Biol.* 14, R183–R184
8 Garcia-Fernàndez, J. (2005) Hox, ParaHox, ProtoHox: facts and guesses. *Heredity* 94, 145–152
9 Santini, S. *et al*. (2003) Evolutionary conservation of regulatory elements in vertebrate hox gene clusters. *Genome Res.* 13, 1111–1122
10 Amores, A. *et al*. (1998) Zebrafish hox clusters and vertebrate genome evolution. *Science* 282, 1711–1714
11 Naruse, K. *et al*. (2000) A detailed linkage map of medaka, *Oryzias latipes*: comparative genomics and genome evolution. *Genetics* 154, 1773–1784
12 Ledje, C. *et al*. (2002) Characterization of Hox genes in the bichir, *Polypterus palmas*. *J. Exp. Zool.* 294, 107–111
13 Hoegg, S. *et al*. (2004) Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *J. Mol. Evol.* 59, 190–203
14 Fried, C. *et al*. (2004) Exclusion of repetitive DNA elements from gnathostome Hox clusters. *J. Exp. Zoolog. B Mol. Dev. Evol.* 302, 165–173
15 Chiu, C-h. *et al*. (2002) Molecular evolution of the *Hox*A cluster in the three major gnathostome lineages. *Proc. Natl. Acad. Sci. U. S. A.* 99, 5492–5497
16 Prohaska, S.J. *et al*. (2004) Surveying phylogenetic footprints in large gene clusters: applications to Hox cluster duplications. *Mol. Phylogenet. Evol.* 31, 581–604
17 Prohaska, S.J. *et al*. (2004) The shark *Hox*N cluster is homologous to the human *Hox*D cluster. *J. Mol. Evol.* 58, 212–217
18 Wagner, G.P. *et al*. (2004) Divergence of conserved non-coding sequences: rate estimates and relative rate tests. *Mol. Biol. Evol.* 21, 2116–2121
19 Wagner, G.P. *et al.* Molecular evolution of duplicated ray-finned fish *Hox*A clusters: increased synonymous substitution rate and asymmetrical co-divergence of coding and non-coding sequences. *J. Mol. Evol.* (in press)
20 Tagle, D.A. *et al*. (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.* 203, 439–455
21 Chiu, C-h. *et al*. (2002) Molecular evolution of the *Hox*A cluster in the three major gnathostome lineages. *Proc. Natl. Acad. Sci. U. S. A.* 99, 5492–5497
22 Loots, G.G. *et al*. (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* 288, 136–140

0168-9525/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2005.06.004

Discovering functional relationships: biochemistry versus genetics

Sharyl L. Wong, Lan V. Zhang and Frederick P. Roth

Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 250 Longwood Ave, Boston, MA, 02115 USA

Biochemists and geneticists, represented by Doug and Bill in classic essays, have long debated the merits of their methods. We revisited this issue using genomic data from the budding yeast, *Saccharomyces cerevisiae*, and found that genetic interactions outperformed protein interactions in predicting functional relationships between genes. However, when combined, these interaction types yielded superior performance, convincing Doug and Bill to call a truce.

Introduction

For more than ten years, Doug, a retired biochemist, and Bill, a retired geneticist, have lived on a hill overlooking a car factory, debating their strategies for reverse engineering a car (see: http://www2.biology.ualberta.ca/locke.hp/dougandbill.htm). Doug advocated rolling up his sleeves, getting under the hood and determining how the parts fit together. Bill preferred tying the hands of a different car-factory worker each morning, then relaxing with a cup of coffee and later examining the cars that emerged from the factory.

One day, Doug and Bill strolled over the next hill. In the midst of debate, they encountered Sharyl, a graduate student in computational genomics. Having overheard their debate, she interjected, 'I don't know much about cars, but I detect an analogy to biochemistry and genetics. I'm trying to discover functional relationships between genes and proteins in yeast and I wonder which of your strategies would work best.'

Corresponding author: Roth, F.P. (fritz_roth@hms.harvard.edu).
Available online 27 June 2005

Differing approaches to determining gene function

To discover functional relationships, Doug would ask, 'Which proteins physically interact with my favorite protein?' By contrast, Bill would perturb the DNA sequence of a gene and observe the consequences *in vivo*, asking 'What are the genetic interaction partners of my favorite gene?' In other words, 'Which genes produce surprising phenotypes if mutated in combination with my favorite gene?' Sharyl described how the fields of biochemistry and genetics had 'gone genomic,' scaling up their classical approaches to discover functional relationships with ever-greater efficiency. Their resulting systematic studies offered a playing field on which to assess Doug and Bill's dilemma. Sharyl then wondered, 'Which type of interaction – protein or genetic – is better at revealing functional relationships?' She pulled out her laptop computer and set to work (Figure 1).

Protein versus genetic interactions in predicting functional relationships

Because 'gene function' is vaguely defined, Sharyl used the Gene Ontology (GO) vocabulary, which describes gene products in terms of biological process, cellular component and molecular function (http://www.geneontology.org/) [1,2]. She defined three measures of functional relatedness for a pair of genes: (i) shared GO biological process (shared process); (ii) shared GO cellular component (shared component); and (iii) shared GO molecular function (shared function). For example, if two genes were assigned to the same GO biological process category, Sharyl considered the gene pair to have a 'shared process'. To avoid associations between genes in broadly defined categories, she considered only specific GO categories – those to which 200 or fewer genes (out of ~6000 total yeast genes) were assigned, including genes assigned to more specific daughter categories. To represent the biochemists, she chose a high-confidence protein-interaction data set based on affinity purification followed by mass spectrometry (APMS) [3]. For the geneticists, she fielded a recent systematic genetic-interaction data set [4] (Tables 1 and 2 in the supplementary data online; Box 1).

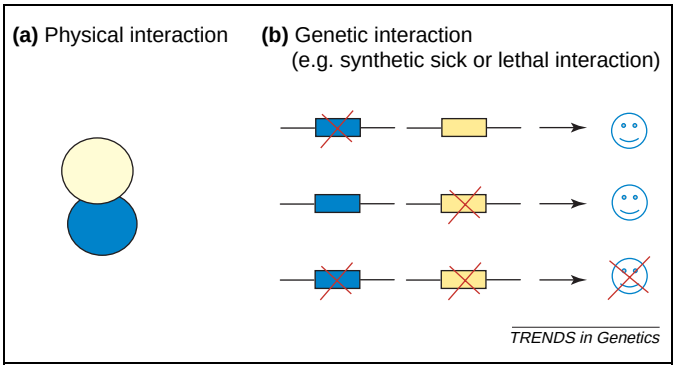

Figure 1. Protein interaction versus genetic interaction. **(a)** A protein interaction exists when two proteins are in physical contact, either direct or indirect (e.g. within the same protein complex). **(b)** By contrast, a genetic interaction is determined between two genes by comparing their single-mutant phenotypes with their double-mutant phenotypes. Here, we focus on synthetic sick or lethal genetic interactions, in which mutation of two genes causes a more severe growth defect (represented by the face marked with an X) than mutation of either alone (represented by happy faces) [4,19]. Yellow and blue circles represent proteins and rectangles represent genes. Rectangles marked with an X represent mutated genes.

To level the playing field, she considered only the 104 409 gene pairs (the 'arena') assessed by both approaches and for which both genes in each pair had a GO annotation. In this arena, the number of gene pairs sharing a specific GO process, component or function was 3841, 1803 and 1139, respectively. The arena contained 48 biochemical interactions and 729 genetic interactions, derived primarily from screens involving the 17 genes used both as baits in the protein-interaction screens and as query genes crossed to 4500 mutants in synthetic genetic array (SGA) analysis. Interestingly, there was no overlap between the protein and genetic interactions (Table 3, supplementary data online). A previous related study [5] did not consider whether gene pairs had been assessed for both types of interaction and used literature-derived interaction data, which are subject to inspection bias.

With a few taps on her keyboard, Sharyl let the games begin. Two proteins exhibiting a protein interaction had a shared process, component or function 42% ($P=2e\text{-}17$), 31% ($P=2e\text{-}15$) and 29% ($P=1e\text{-}16$) of the time, respectively. Genetic interactions were uniformly less-accurate indicators of shared function, with corresponding rates of

Box 1. Protein and genetic-interaction screens

- Synthetic genetic array (SGA) analysis is a high-throughput method that assesses pairs of genes for genetic interaction [4,19]. A strain carrying a mutated query gene is crossed to an array of ~4700 strains, each mutated in a different non-essential yeast gene. The resulting double mutants are then assessed for fitness. Slow growth or lethality relative to each of the single-mutant strains is declared synthetic sickness or lethality. In the SGA data set used here, 159 query genes were crossed to the array, resulting in ~730 000 gene pairs tested for genetic interaction. Based on this data set, the genetic network is between two and 54 times more dense than the protein-interaction network.
- Affinity purification followed by mass spectrometry (APMS) is used for high-throughput discovery of physical protein interactions. A 'bait' protein is precipitated in a complex with its interacting proteins. Members of this 'pulled-down' complex are then identified by mass spectrometry. The two large APMS studies in yeast are known as the tandem affinity purification (TAP) [3] and high-throughput mass spectrometric protein complex identification (HMS-PCI) [6] studies. In both studies, the data can be interpreted in two ways. The spoke interpretation defines an interaction between a bait protein and each protein it pulls down. The matrix interpretation, however, counts interactions between all pairs of proteins pulled down by a bait. In the TAP study, bait constructs were integrated into the yeast genome and expression was controlled by an endogenous promoter. In the HMS-PCI study, however, the bait construct was plamid-borne and expression was controlled by a robust exogenous promoter. Thus, the TAP data set is more likely to be physiologically relevant, although the HMS-PCI study could detect interactions between gene products not normally expressed in the condition examined. The TAP and HMS-PCI data sets employed 1167 and 725 baits, respectively. A gene pair was considered assessed for protein interaction, if at least one gene of the pair was a bait and the other was not filtered out as a 'promiscuous prey' [6].
- Yeast-two-hybrid (Y2H) is a high-throughput method for assessing direct physical interaction between two proteins (although indirect 'bridged' interactions can also be detected). Here our Y2H data set consisted of the union of the interactions reported by Uetz *et al.* [18] and the 'core' version (corresponding to interactions detected at least three times) of the data set produced by Ito *et al.* [17].

Glossary

Accuracy: is defined as the number of gene pairs with the same function divided by the number of gene pairs with the given predictive characteristic. For example, the number of pairs that both genetically interact and have shared process divided by the number of pairs that genetically interact.
Sensitivity (or true positive rate): is defined as the number of gene pairs with the same function that pass a given score threshold (i.e. true positives) divided by the total number of gene pairs with the same function.
False positive rate (or 1 – specificity): is defined as the number of gene pairs without the same function that pass a given score threshold (i.e. false positives) divided by the total number of gene pairs without the same function.

19% (P=2e-63), 15% (P=2e-66) and 8% (P=2e-28). However, genetic interactions detected gene pairs with shared function with much higher sensitivity (4–6%) than biochemical interactions (0.5–1.2%; Table 4 in the supplementary data online). When considering different physical-interaction data sets [3,6] (Box 1), genetic interactions were consistently more sensitive and sometimes more accurate (see Glossary; Table 4, supplementary data online). Thus, it was difficult to declare a clear winner.

Combining genetic and protein interactions with other data

Are genetic interactions combined with other types of evidence more informative than protein interactions combined with other evidence? Rather than considering each type of interaction in isolation, several groups have previously combined heterogeneous data, using machine learning approaches to predict some property of a gene pair or to predict gene function [7–12]. Therefore, Sharyl combined multiple types of evidence [11] – including co-localization [13], sequence homology [14], correlated mRNA expression [15,16] and chromosomal distance (Table 5, supplementary data online) – to predict shared function. She chose a previously described probabilistic-decision tree approach [12] and compared performance with and without the benefit of protein and/or genetic-interaction data. For each of shared process, component, and function and for each choice of input data, she performed cross-validation: she randomized all gene pairs in the arena into four groups, and successively scored each group using a model trained on the remaining three. She then compared the prediction score of each gene pair with its corresponding shared process, function or component status. A plot of true- versus false-positive rates revealed that genetic and protein interactions were comparable at low sensitivities; however, as sensitivity increased, genetic-interaction data enhanced performance more than protein-interaction data. This trend was observed for shared process (Figure 2), component (Figure 1a, supplementary data online) and function (Figure 1b, supplementary data online). Doug, the biochemist, began to despair.

Before Bill could begin to gloat, however, Sharyl showed that genetic- and protein-interaction data together gave markedly better results than either alone, suggesting that each offers distinctly different types of information. Although protein interactions can represent associations between genes in the same complex or physically connected pathway, genetic interactions can additionally reflect relationships between genes in physically non-interacting pathways. She repeated this analysis with another APMS protein-interaction data set [6] and then with the union of two yeast-two-hybrid (Y2H) data sets [17,18] (Tables 1 and 3, and Figures 2 and 3 in the supplementary data online), altering the arena appropriately. In each case, genetics beat biochemistry by a slim margin, but the combination of these complementary interaction types outperformed either alone. Sharyl's results convinced Doug and Bill to shake hands and head back over the hill … until new data or new technology call for a rematch.

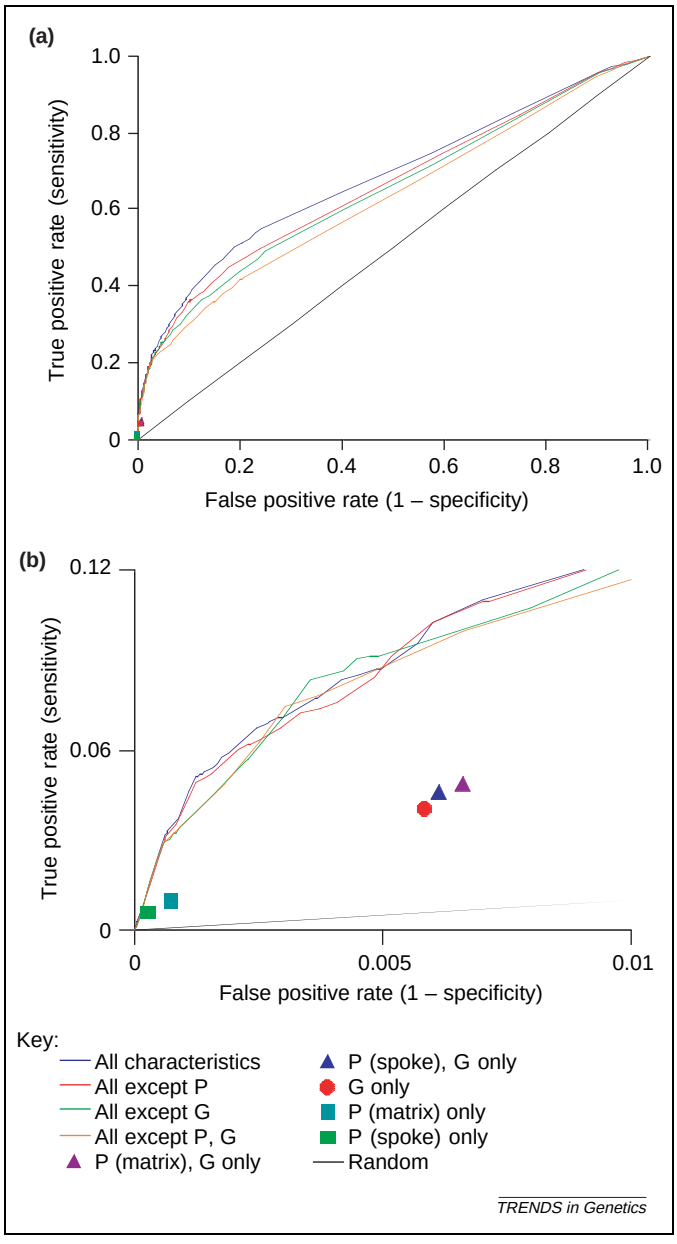

Figure 2. Performance when predicting 'shared process' with and without genetic [4] and/or protein-interaction data from Gavin *et al.* [3]. **(a)** Each point on the curve represents performance at a given threshold score (such that pairs above that threshold are predicted to have shared function). 'P' and 'G' represent protein and genetic-interaction data respectively. Using all high-throughput data yields the best prediction performance. Notably, this performance is impaired more by the omission of genetic interaction data than by omission of protein interaction data. The same information is shown in **(b)** but in finer detail.

Acknowledgements

We thank C. Boone, H. Fraser, and T. Hughes for helpful comments, O. King for mathematical advice and G. Berriz for his Gene Ontology parser. S.L.W. was supported by the Ryan Foundation and the Milton Fund of Harvard University. L.V.Z. was supported by the Fu Foundation, the Ryan Foundation and the American Association of University Women. This work was also supported by Funds for Discovery provided by John Taplin, a Howard Hughes Medical Institute institutional grant to Harvard Medical School and the National Institutes of Health/National Human Genome Research Institute.

Supplementary data

Supplementary data associated with this article can be found at doi:10.1016/j.tig.2005.06.006

References

1 Harris, M.A. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32 Database issue, D258–261
2 Dwight, S.S. *et al*. (2004) *Saccharomyces* genome database: underlying principles and organisation. *Brief. Bioinform.* 5, 9–22
3 Gavin, A.C. *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141–147
4 Tong, A.H. *et al*. (2004) Global mapping of the yeast genetic interaction network. *Science* 303, 808–813
5 Deng, M. *et al*. (2004) An integrated probabilistic model for functional prediction of proteins. *J. Comput. Biol.* 11, 463–475
6 Ho, Y. *et al*. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415, 180–183
7 Troyanskaya, O.G. *et al*. (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl. Acad. Sci. U. S. A.* 100, 8348–8353
8 Lee, I. *et al*. (2004) A probabilistic functional network of yeast genes. *Science* 306, 1555–1558
9 Chen, Y. and Xu, D. (2004) Global protein function annotation through mining genome-scale data in yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res.* 32, 6414–6424
10 Zhang, L.V. *et al*. (2004) Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics* 5, 38
11 Wong, S.L. *et al*. (2004) Combining biological networks to predict genetic interactions. *Proc. Natl. Acad. Sci. U. S. A.* 101, 15682–15687
12 King, O.D. *et al*. (2003) Predicting gene function from patterns of annotation. *Genome Res.* 13, 896–904
13 Kumar, A. *et al*. (2002) Subcellular localization of the yeast proteome. *Genes Dev.* 16, 707–719
14 Altschul, S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
15 Hughes, T.R. *et al*. (2000) Functional discovery via a compendium of expression profiles. *Cell* 102, 109–126
16 Cho, R.J. *et al*. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* 2, 65–73
17 Ito, T. *et al*. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U. S. A.* 98, 4569–4574
18 Uetz, P. *et al*. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 623–627
19 Tong, A.H. *et al*. (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* 294, 2364–2368

0168-9525/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2005.06.006