Chapter 20

Effects of Functional Bias on Supervised Learning of a Gene Network Model

Insuk Lee and Edward M. Marcotte

Abstract

Gene networks have proven to be an effective approach for modeling cellular systems, capable of capturing some of the extreme complexity of cells in a formal theoretical framework. Not surprisingly, this complexity, combined with our still-limited amount of experimental data measuring the genes and their interactions, makes the reconstruction of gene networks difficult. One powerful strategy has been to analyze functional genomics data using supervised learning of network relationships based upon reference examples from our current knowledge. However, this reliance on the set of reference examples for the supervised learning can introduce major pitfalls, with misleading reference sets resulting in suboptimal learning. There are three requirements for an effective reference set: comprehensiveness, reliability, and freedom from bias. Perhaps not too surprisingly, our current knowledge about gene function is highly biased toward several specific biological functions, such as protein synthesis. This functional bias in the reference set, especially combined with the corresponding functional bias in data sets, induces biased learning that can, in turn, lead to false positive biological discoveries, as we show here for the yeast *Saccharomyces cerevisiae*. This suggests that careful use of current knowledge and genomics data is required for successful gene network modeling using the supervised learning approach. We provide guidance for better use of these data in learning gene networks.

Key words: Gene network model, supervised learning, classification, functional coupling, functional bias, reference set, genomics data.

1. Introduction

A major goal for our system-level understanding of a cell or an organism is the identification of the functions of all genes/proteins and their organization into pathways. With the classical one-gene-one-study approach, this goal is certainly daunting, if not impossible. However, the massive generation of biological data by

Jason McDermott et al. (eds.), *Computational Systems Biology, vol. 541*

DOI 10.1007/978-1-59745-243-4_20

high-throughput techniques developed over the past decade brings this ambitious goal much closer, and abundant functional genomics data provide opportunities for modeling global gene/protein networks, which may shed light on our understanding of cellular systems.

Supervised machine-learning approaches have recently become popular in global gene/protein network modeling for various organisms *(1–5)*. Supervised learning is generally considered a "classification" task, in which we start with classes predefined by some criterion (usually given by expert opinion) and attempt to find additional cases of these from the data. For modeling gene networks, the typical approach is not the prediction of genes with a completely defined set of cellular functions – this strategy is difficult, not least, because the total set of gene functions is unknown and because many gene functions overlap. Instead, networks are often derived by examining two classes of gene pairs, functionally coupled or not. Note that the network models are intrinsically consistent with genes' pleiotropic (multi-functional) natures. Connections (perhaps weighted) within such networks capture functional relationships among genes and can therefore be used to discover functions of uncharacterized genes, to define functional modules of genes, and to describe the organization of genes that contribute to the physiological state of the cell.

Learning by classification requires reference examples on which to train, and in this case they would be known, functionally coupled gene pairs. A set of reference examples is generally based on current knowledge and expert opinion. Reliable examples of gene functional coupling can be derived easily from various biological annotation sets based mostly on manual curation by expert biologists (**Table 20.1**). In order to allow effective supervised learning, a reference set must

Table 20.1
Annotation databases for gene functions

GO (gene ontology) biological process http://www.geneontology.org/ontology/process.ontology GO is hierarchically organized, with the top-level (level 0) annotation being most general and the bottom level the most specific. Generally, the middle range of annotation provides a good compromise between specificity and comprehensiveness.
Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway ftp://ftp.genome.jp/pub/kegg/pathways/sce/sce_gene_map.tab KEGG offers a three-level hierarchical annotation of biological pathways. The bottom-level terms are most useful as functional reference terms, but show a bias toward metabolic pathways.
CYGD (the comprehensive yeast genome database) functional category, hosted by MIPS ftp://ftpmips.gsf.de/yeast/catalogues/funcat/ CYGD is a reasonably comprehensive and detailed annotation set that is specific for yeast. The top level contains 11 broad functional categories that are useful for visualization and analysis of general functional trends.

clearly be both comprehensive and reliable. However, reference sets for functional networks have another important requirement, which is freedom from functional bias. In fact, current biological annotations often display severe bias toward a few specific functions. In this chapter, we demonstrate how this reference set bias affects the investigation of functional linkages from diverse genomics data, and we present examples for the genes of the yeast *S. cerevisiae*.

2. Methods

2.1. Functional Bias in Current Functional Annotations

A number of different databases organize genes according to their pathways, such as the three listed in **Table 20.1**. Among these, gene ontology (GO) annotation has become popular for functional genomics studies due to its hierarchical organization and its separation of three aspects of gene function—biological process, which captures pathway relationships; cellular component, which describes sub-cellular localization of gene products; and molecular function, which focuses more on enzymatic and binding functionalities *(6)*. GO has also consistently improved through community efforts *(7)*. For example, by March 2005, 4,199 yeast genes (~72% of the total of 5,794 verified protein encoding genes) were annotated by at least one GO biological process annotation. Therefore, a functional annotation reference set based on GO biological processes is highly comprehensive, satisfying the first requirement for effective supervised learning.

Another requirement for an effective reference set is reliability. We can control the reliability of GO-derived reference sets both at the level of their generality and with regard to the evidence supporting them. First, we can control the generality of employed—general annotations such as metabolism (GO:0008152) are typically located near the top of the GO hierarchy and often provide poor resolution in the learning of specific cellular functions. By contrast, annotations near the bottom of the GO hierarchy are highly specific but annotate only one or a few genes, and thus they lack comprehensiveness. The middle layers of the gene ontology hierarchy generally provide a more optimal trade-off between comprehensiveness and reliability (*see* **Note 1**). GO also provides evidence codes (**Table 20.2**) as another way of controlling the reliability of the reference set. Annotations by traceable author statement (TAS), inferred from direct assay (IDA), inferred from mutant phenotype (IMP), inferred from genetic interaction (IGI), and inferred from physical interaction (IPI) are generally considered as highly reliable annotations. As of March, 2005, the GO biological process had 4,199 annotated yeast genes with a total 11,430 terms, of which 9,093 (~80%) are based on one of these five types of highly reliable evidence.

Table 20.2
Gene ontology evidence codes and their reliability

Code	Description	Reliability
TAS	Traceable Author Statement	High
IDA	Inferred from Direct Assay	High
IMP	Inferred from Mutant Phenotype	High
IGI	Inferred from Genetic Interaction	High
IPI	Inferred from Physical Interaction	High
ISS	Inferred from Sequence or Structural Similarity	Low
IEP	Inferred from Expression Pattern	Low
NAS	Non-traceable Author Statement	Low
IEA	Inferred from Electronic Annotation	Low

The reference set obtained by considering the above GO annotations, though highly reliable and comprehensive, may still have a systematic bias toward a few specific functions. This in turn may lead to biased learning of the given genomics data. We examined the distribution of GO biological process terms from the middle layers of the annotation hierarchy (between levels 6 and 10, *see* **Note 1**). Although we expected similar genome coverage among functional terms, we found a few dominant functional terms used to annotate yeast genes. From 1,067 selected GO biological process terms, we observed that a single functional term, protein biosynthesis (GO:0006412), accounts for more than 4% of total gene annotations, although its expected coverage is less than 0.1% ($100 / 1{,}067 < 0.1$) (**Fig. 20.1A**, filled bars).

In order to evaluate functional coupling between genes, we derived reference gene pairs that are functionally coupled (positives) and pairs that are not functionally coupled (negatives) from the given gene functional annotation set. The simplest way of deriving a set of positive examples is to pair genes that share at least one common functional description. A corresponding set of negative examples can be derived by pairing genes that do not share any functional description (*see* **Note 2**). As a result of this gene pairing, the functional bias of gene annotation is dramatically amplified in the reference sets of functionally coupled gene pairs (**Fig. 20.1A**, empty bars). This functional bias annotation is not specific to a particular annotation set. We observe a similar functional bias toward the ribosome, the core machinery of protein biosynthesis, in another commonly used gene function annotation set, the Kyoto Encyclopedia of Genes and Genomes (KEGG) *(8)*, which focuses mainly on metabolic pathway information (**Fig. 20.1B**).

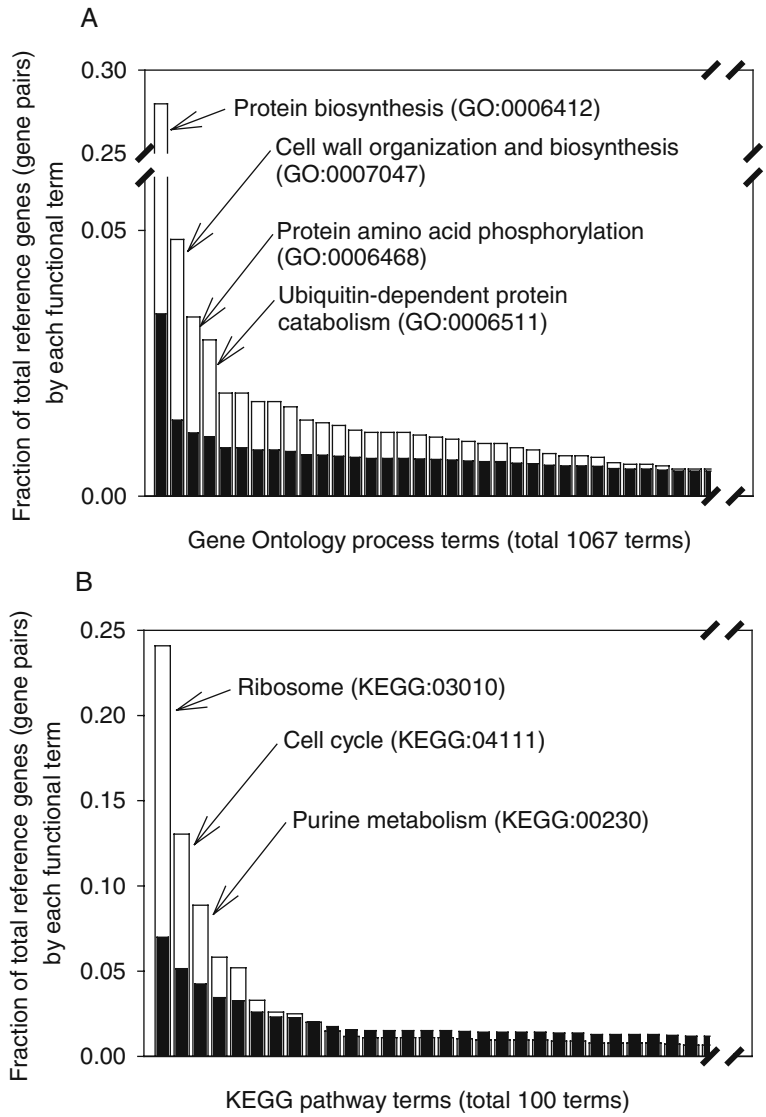

Fig. 20.1. The distributions of functional terms among annotated genes (*filled bars*) and gene pairs (*empty bars*) by (**A**) the gene ontology (GO) biological process annotations, and (**B**) the Kyoto Encyclopedia of Genes and Genomes (KEGG). The pathways illustrate functional bias in both major annotation databases. Only a few of the most dominant terms out of the 1,067 GO biological process terms between levels 6 and 10 and out of 100 KEGG pathway terms at the bottom level are labeled. In both GO biological process and KEGG, the most dominant functional term is related to protein biosynthesis (the ribosome being the major component represented), and this single term accounts for 3 and 7% of total gene annotations by GO and KEGG, respectively. This functional bias in gene annotations has become dramatically amplified by pairing genes for the same functional terms to provide references of gene pairs functionally coupled. As references of gene pairs, about 25% of total reference examples are based upon only a single most dominant functional term in both annotation sets.

There are two major systematic biases that introduce functional biases into our current knowledge bases. First, there are differences in the size of cellular functional modules. We will refer to this as size bias. For example, the protein biosynthesis functional module consists of a huge translational machine, the ribosome (composed of 150 $\sim$ 200 proteins in yeast), as well as many other co-factors. Thus, this single specific functional module annotates many more genes than any other. The second bias we will refer to as a study bias. Biologists have historically studied genes using a one-gene-one-study approach, which naturally introduced a bias toward genes that are more important (and often biologically more essential) or those more readily studied (for example, genes with an obvious mutation phenotype are easier to study). Genes involved in protein biosynthesis have been subject to this study bias. The recent development of various high-throughput functional genomics analyses with reverse-genetics approaches solves the problem of study bias, but size bias is an intrinsic characteristic of cellular systems.

2.2. Effect of Reference Set Functional Bias on Supervised Learning

The inevitable functional bias in gene annotations potentially affects further discovery of gene functions and network organization. To demonstrate the effect of a single dominant functional term (protein biosynthesis, *see* Fig, **20.1A**) of the reference set based on the GO biological process in learning functional gene coupling, we compare two different reference sets derived from the GO biological process annotation set: (1) a *biased reference set* that comprises all gene pairs sharing annotation, including the pairs sharing the function "protein biosynthesis"; and (2) an *unbiased reference set* based on the same pairs but excluding those sharing the protein biosynthesis term. The effects on learning for links are illustrated in **Fig. 20.2** with examples from various types of yeast genome-wide functional genomics data.

We can infer which genes are functionally coupled by the co-expression patterns across different experimental conditions. The tendency toward co-expression can be measured by the Pearson correlation coefficient between any two genes' expression profile vectors. For a set of gene pairs with a given range of co-expression tendencies, we calculate the log-likelihood score (*LLS*) using Bayesian statistics, as a measurement of the likelihood of functional coupling supported by the given data (*see* **Note 3**). In **Fig. 20.2**, log-likelihood scores are calculated with the 0.632 bootstrapping method *(9)* to minimize the over-fitting of models (*see* **Note 4**). For the functionally informative microarray data set, we observe a significant positive correlation between the tendency toward co-expression and the measured likelihood of functional coupling between pairs of genes.

For microarray data from yeast cell cycle time courses, log-likelihood scores are higher with the biased reference set than with the unbiased one (**Fig. 20.2A**). For the most significant data range

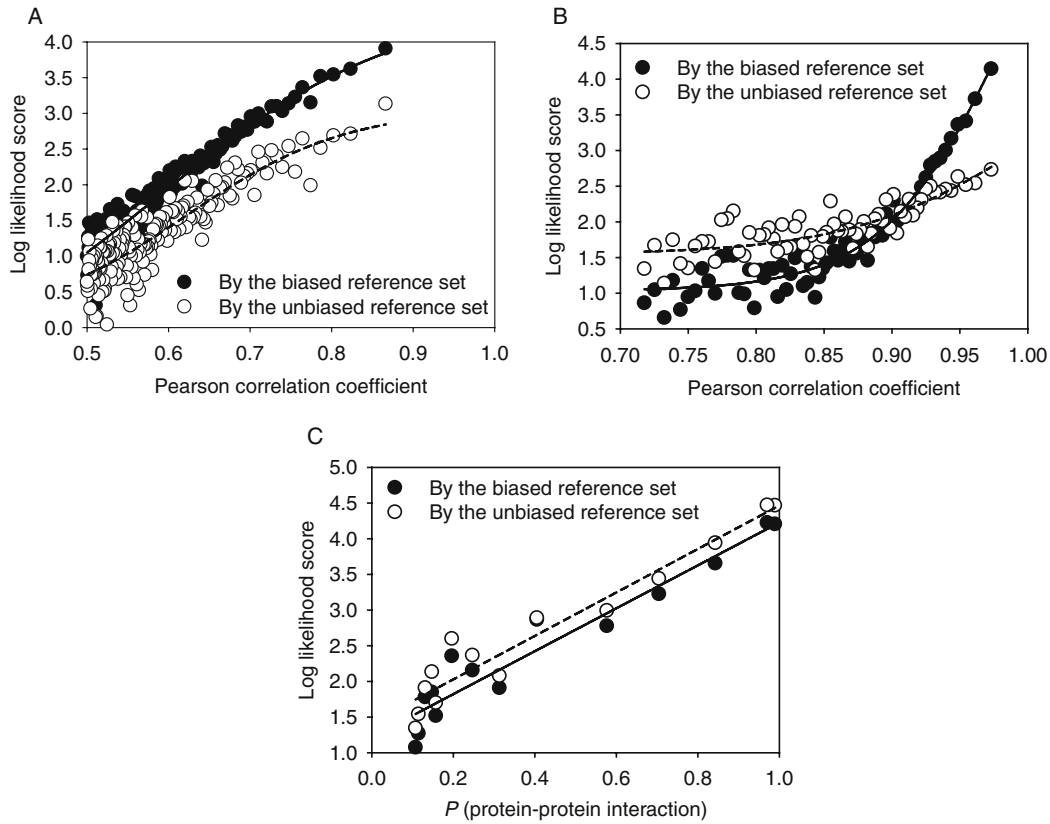

Fig. 20.2. Correlation between data-intrinsic scores that imply gene functional coupling (Pearson correlation coefficient measuring co-expression tendency or probability score of protein–protein interaction) and log-likelihood score (*see* **Note 3**) that measures the likelihood of gene functional coupling with the given supporting data. Three different data sets—(**A**) microarray data with cell cycle time courses, (**B**) microarray data with various heat-shock conditions, and (**C**) protein–protein interaction from affinity complex purification—are evaluated using two different reference sets derived from GO biological process annotation: (1) the biased set (*filled circle*), including gene pairs among the most dominant term "protein biosynthesis," and (2) the unbiased set (*empty circle*), excluding reference gene pairs for the term "protein biosynthesis." Each data point represents a bin of 1,000 gene pairs of the data set, which are sorted by data-intrinsic scores.

(the top 1,000 gene pairs), the log-likelihood score is about 4 (i.e., the likelihood is ~55-fold higher than random) with the biased reference set, while it is about 3 (i.e., the likelihood is ~20-fold higher than random) with the unbiased reference set. Thus, there is an increase of approximately threefold overall, deriving entirely from a single additional functional term. This observation becomes extreme as we examine the microarray data collected from heat-shock-treated cells (**Fig. 20.2B**). The positive correlation between the co-expression of genes across heat-shock conditions, and the likelihood of functional coupling, is very strong with the biased reference set, especially when the Pearson correlation coefficient is higher than 0.8. In the range of Pearson correlation coefficients from 0.8 to 1, the *LLS* increases from ~1 (i.e., the

likelihood is approximately threefold higher than random) to ~ 4.2 (likelihood ~66-fold higher than random), achieving a ~22-fold increase in the likelihood of functional coupling for the most co-expressed genes, apparently implying that this heat-shock microarray data carries strong information about gene functional couplings. However, masking the single dominant functional term, protein biosynthesis, is sufficient to remove most of the trend, showing only approximately threefold likelihood increase across the same range of Pearson correlation coefficients (from *LLS* ~1.7, 5.5-fold higher likelihood than random, to *LLS* ~2.7, 15-fold higher likelihood than random).

This over-optimism exhibited by the biased reference set largely disappears when we consider protein–protein interaction data. We compared the biased and the unbiased reference sets in evaluating a high-throughput protein–protein interaction data set derived from affinity purification of protein complexes followed by mass spectrometry analysis *(10)*. Using machine-learning algorithms, the raw data set has been simplified to a set of 14,317 protein–protein physical interactions with associated probabilistic scores *(10)*. In this data set, the biased reference actually provides very similar likelihood values to the unbiased reference (**Fig. 20.2C**). A similar trend is evident in a high-quality data set of 12,300 interactions derived from published protein physical and genetic interaction data (and excluding large-scale assay-derived interaction data) *(11)*. This data set shows a very high overall quality and relatively little difference in performance between the unbiased reference set ($LLS = 3.85$) and the biased reference set ($LLS = 3.55$).

2.3. Effect of Genomics Data Set Functional Bias on Supervised Learning

What are the underlying characteristics of data sets sensitive to this reference set bias? Not surprisingly, in many data sets these appear to be functional biases that affect their performance in supervised learning. This trend is evident when measuring the functional bias as a function of interaction confidence score (the gene retrieval rate). The gene retrieval rates measured for genes of 11 different functional groups defined by the Munich Information Center for Protein Sequences (MIPS) *(12)* demonstrates a high bias toward genes involved in protein biosynthesis, which are among the most highly co-expressed gene pairs in yeast cell-cycle microarray experiments (**Fig. 20.3A**). This trend explains the overly optimistic evaluation of cell-cycle micro-array data sets by the biased reference set. It has been shown that proteins in stable complexes tend strongly to co-express *(13)*. Therefore, co-expression of genes is an excellent feature for inferring interactions among proteins of stable complexes such as the ribosome, and, not surprisingly, the most strongly co-expressing genes are highly enriched for ribosomal protein pairs. This trend is exacerbated through the use of the biased reference set, which is over-represented for

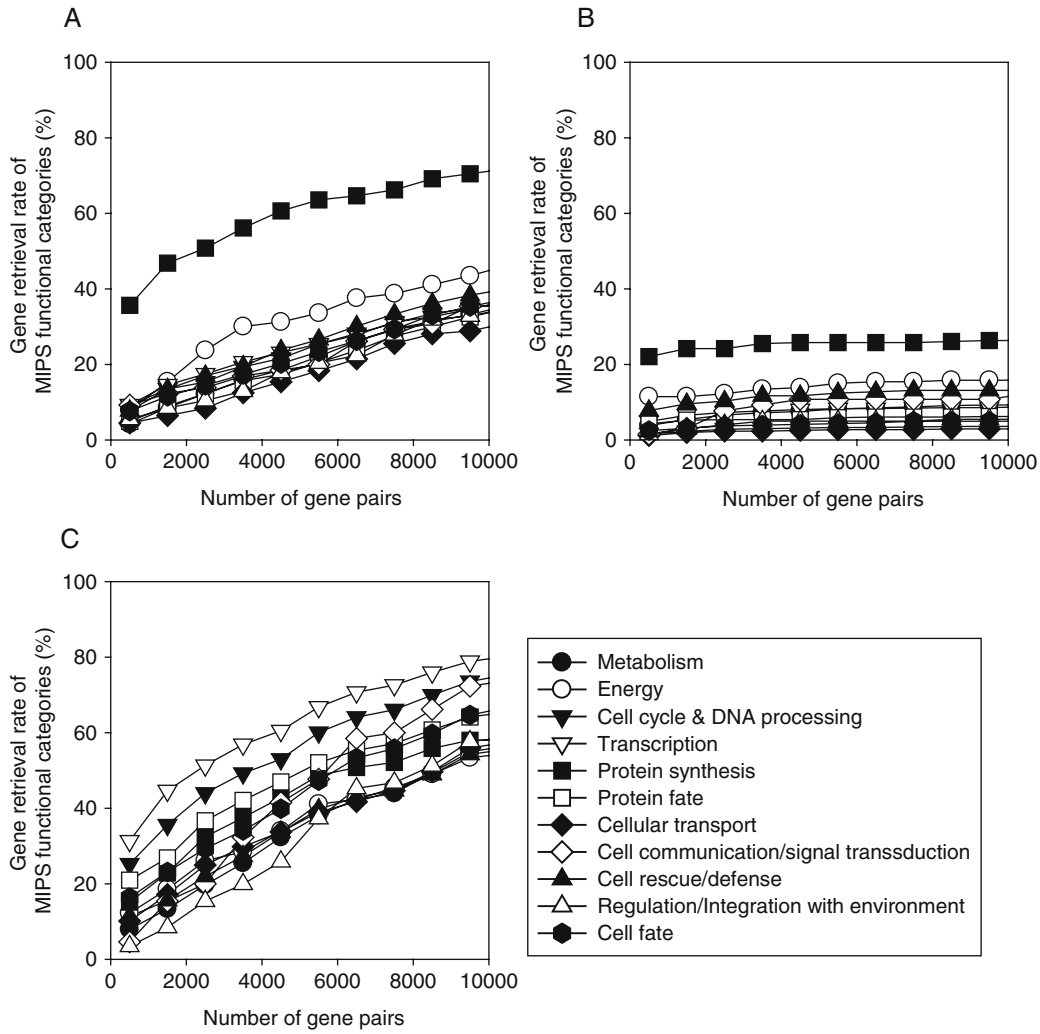

Fig. 20.3. Gene retrieval rate across 11 MIPS functional categories for the top 10,000 scored functional gene pairs in (**A**) co-expression during cell cycle time courses, (**B**) co-expression across various heat-shock conditions, and (**C**) protein–protein interactions from affinity complex purifications. The cumulative gene functional coverage for the given data set was measured based on MIPS' 11 top-level protein functional categories for every 1,000 gene pairs sorted by data-intrinsic scores.

ribosomal gene pairs. Thus, the generality of interactions discovered from these data may be suspect when the biased reference set is used.

The distribution of gene functions in the yeast heat-shock microarray data set is also interesting. While showing a similar enrichment of protein biosynthesis genes, this set also shows a flat gene retrieval rate (**Fig. 20.3B**)—i.e., the gene retrieval rate does not significantly increase with an increasing number of co-expressed gene pairs. This implies that co-expression during yeast heat shock is restricted to only a small percentage of cellular

systems. In this case, additional functional couplings are added to this small set of systems (which include the ribosome and other protein biosynthesis-related systems) without incorporating additional genes, leading to an increasingly dense network of functional linkages among this subset of genes. This is evident in the fact that functional couplings derived from co-expression on the heat-shock data cover only 9% of the yeast proteome and have a clustering coefficient *(14)* of 0.6 for the top 10,000 gene pairs, implying highly clustered interactions in a relatively few functional modules. By contrast, the same number of gene pairs derived from the cell-cycle data set covers about 36% of the proteome and has a lower clustering coefficient (0.28). The protein–protein interaction data sets show less functional bias (**Fig. 20.3C**), with large proteome coverage (58%) and a low clustering coefficient (0.14) for the number of gene pairs, implying that the information in these sets is distributed through many functional modules in the yeast cell.

2.4. Circumventing Functional Bias in Reference and Data Sets

How can we achieve reliable evaluation in the presence of persistent functional bias in the reference and data sets? Approaches for monitoring over-training, such as cross-validation and bootstrapping, do not solve this problem, as seen in the results of **Fig. 20.2**, which were carried out with 0.632 bootstrapping (*see* **Note 4**). One simple approach is to ignore the dominant terms for the purposes of training and testing. For an unbiased data set, this masking of a dominant functional term has minimal effects, as we show with the example of protein–protein interaction data (**Fig. 20.2C** and **3C**). However, for biased data sets (**Fig. 20.2A**, **B** and **20.3A**, **B**), we observe much lower likelihoods of functional coupling, implying that the optimistic likelihood scores were unrealistic and therefore risky to generalize to the rest of the data. Combined with cross-validation or bootstrapping, this dominant term masking is a simple but effective way to remove much of the negative effects of functional bias toward a few dominant functional annotations.

3. Notes

1. *Hierarchy in gene ontology.* The gene ontology is hierarchically organized, and references derived from different levels of the annotation hierarchy may result in quite different evaluations for identical data sets. This hierarchy is diamond-shaped, characterized by fewer descriptive terms at the top and bottom levels and by a gradual increase in terms as one moves toward

the middle layer of the annotation hierarchy. Generally speaking, top-level annotations provide extensive coverage but low information specificity (resolution), while low-level annotations describe fewer genes but with high specificity. Therefore, the trade-off between annotation coverage and specificity must be considered carefully in order to obtain an effective reference set for evaluating genomics data. Empirically, we find good performance using GO biological process terms between level 6 and 10 out of the total of 15 levels. The term "biological process" is considered to be level 0.

2. *Imbalance between positive and negative examples in a reference set of gene functional couplings.* Generating positive (negative) reference examples of functionally coupled gene pairs by pairing genes sharing (not sharing) any functional annotation results in a serious imbalance in the sizes of the two reference sets. We obtain a much larger negative reference set than positive (e.g., ~100-fold larger negative reference set than positive based on the yeast GO biological process annotation of March 2005). This much higher frequency of negative examples in a reference set is problematic if one uses conventional data evaluation methods, which use a "true positive rate" (true positive / predicted as positive) such as a recall-precision curve (generally an overly pessimistic evaluation indicated by low precision for a given recall) or receiver operating characteristic (ROC) curve (generally an overly optimistic evaluation indicated by a high true positive rate for a given false positive rate) *(15)*. With the severe size imbalance between the positive and negative reference sets, the measurement of the true positive rate is often discouraging in absolute terms; however, as a relative measure among different data sets, it works well. Gene functional couplings can be learned using these relative reliability scores, and various thresholds of scores will generate gene network models with varying accuracies and differing coverage.

3. *Evaluation of gene functional coupling by log likelihood scores.* We can evaluate the reliability of gene functional couplings supported by the given data using Bayesian statistics. A formal representation of Bayesian inference of the functional coupling between genes is the log likelihood score (*LLS*),

$$LLS = \ln\left(\frac{P(I|D)/P(\sim I|D)}{P(I)/P(\sim I)}\right),$$

where $P(I|D)$ and $P(\sim I|D)$ are the frequencies of gene functional coupling and its negation observed in the given genomics dataset (D), as measured by reference gene pairs. $P(I)$ and $P(\sim I)$ represent the prior expectations (the total frequencies of all positive and negative reference gene pairs,

respectively). A score of zero indicates coupled gene pairs in the data being tested are no more likely to be functionally coupled than random; higher scores indicate a more informative data set for identifying functional relationships.

4. *Evaluation with 0.632 bootstrapping.* To avoid over-fitting, we employed 0.632 bootstrapping *(9)* for all *LLS* evaluations. The 0.632 bootstrapping has been shown to provide a robust estimate of functional coupling accuracy. It is especially favored over cross-validation for very small datasets *(9)*. Data evaluation with bootstrapping is therefore appropriate even for more poorly annotated genomes. Unlike cross-validation, which uses multiple tests and training sets by sampling data without replacement, 0.632 bootstrapping constructs the training set from data sampled with replacement and the test set from the non-sampled data. For the sampling, each instance has a probability of $1–1/n$ of not being sampled, resulting in ~63.2% of the data being in the training set and ~36.8% in the test set *(16)*. The overall *LLS* is the weighted average of results for the two sets with 10 repetitions, equal to $0.632*LLS_{\text{test}} + (1–0.632)*LLS_{\text{train}}$.

Acknowledgments

This work was supported by grants from the N.S.F. (IIS-0325116, EIA-0219061, 0241180), N.I.H. (GM06779-01), Welch (F-1515), and a Packard Fellowship (E.M.M.). We thank Cynthia V. Marcotte and Ray Hardesty for help with editing.

References

1. Jansen, R., Yu, H., et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 2003; 302:449–53.
2. Lee, I., Date, S. V., et al. A probabilistic functional network of yeast genes. Science 2004; 306:1555–8.
3. Myers, C. L., Robson, D., et al. Discovery of biological networks from diverse functional genomic data. Genome Biol 2005; 6:R114.
4. Rhodes, D. R., Tomlins, S. A., et al. Probabilistic model of the human protein-protein interaction network. Nat Biotechnol 2005; 23:951–9.
5. Zhong, W., and Sternberg, P. W. Genome-wide prediction of *C. elegans* genetic interactions. Science 2006; 311:1481–4.
6. Ashburner, M., Ball, C. A., et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000; 25:25–9.
7. Cherry, J. M., Adler, C., et al. SGD: Saccharomyces genome database. Nucleic Acids Res 1998; 26:73–9.
8. Kanehisa, M., and Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000; 28:27–30.
9. Efron, B., and Tibshirani, R. An introduction to the bootstrap. New York: Chapman & Hall, 1993.
10. Krogan, N. J., Cagney, G., et al. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. Nature 2006; 440:637–43.

11. Reguly, T., Breitkreutz, A., et al. Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. J Biol 2006; 5:11.
12. Mewes, H. W., Amid, C., et al. MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Res 2004; 32:D41–4.
13. Jansen, R., Greenbaum, D., et al. Relating whole-genome expression data with protein-protein interactions. Genome Res 2002; 12:37–46.
14. Watts, D. J., and Strogatz, S. H. Collective dynamics of 'small-world' networks. Nature 1998; 393:440–2.
15. Jansen, R., and Gerstein, M. Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. Curr Opin Microbiol 2004; 7:535–45.
16. Witten, I. H., and Frank, E. Data Mining: Practical Machine Learning Tools and Techniques. San Francisco, CA: Morgan Kaufmann, 2005.