Newcastle University

COMPUTING SCIENCE

Multiple Gold Standards Address Bias in Functional Network Integration

Katherine James, Samantha J. Lycett, Anil Wipat and Jennifer S. Hallinan

TECHNICAL REPORT SERIES

No. CS-TR-1302 November 2011

TECHNICAL REPORT SERIES

No. CS-TR-1302 November, 2011

Multiple Gold Standards Address Bias in Functional Network Integration

K. James, S.J. Lycett, A. Wipat and J.S. Hallinan

Abstract

Network integration is a widely-used method of combining large, diverse data sets. Edge weights, representing the probability that an edge actually exists, can add greatly to the value of the networks. The edge weights are usually calculated using a Gold Standard dataset. However, all Gold Standards suffer from incomplete coverage of the genome, and from bias in the type of interactions detected by different experimental techniques. Consequently the use of a single Gold Standard tends to bias the integrated network. We describe a novel Bayesian Data Fusion method for selecting and using multiple Gold Standards for scoring datasets prior to integration. We demonstrate the utility of networks scored against multiple Gold Standards for the pre-diction of Gene Ontology annotations for genes from KEGG pathways. Finally, we apply the networks to the functional prediction of genes which were uncharacterised in datasets from 2007, and evaluate the network results in the light of recent annotations.

Printed and published by Newcastle University,
Computing Science, Claremont Tower, Claremont Road,
Newcastle upon Tyne, NE1 7RU, England.

Bibliographical details

HAO, F., KREEGER, M.N.

Multiple Gold Standards Address Bias in Functional Network Integration
[By] K. James, S.J. Lycett, A. Wipat, J.S. Hallinan
Newcastle upon Tyne: Newcastle University: Computing Science, 2011.

(Newcastle University, Computing Science, Technical Report Series, No. CS-TR-1302)

Added entries

NEWCASTLE UNIVERSITY
Computing Science. Technical Report Series. CS-TR-1302

Abstract

Network integration is a widely-used method of combining large, diverse data sets. Edge weights, representing the probability that an edge actually exists, can add greatly to the value of the networks. The edge weights are usually calculated using a Gold Standard dataset. However, all Gold Standards suffer from incomplete coverage of the genome, and from bias in the type of interactions detected by different experimental techniques. Consequently the use of a single Gold Standard tends to bias the integrated network. We describe a novel Bayesian Data Fusion method for selecting and using multiple Gold Standards for scoring datasets prior to integration. We demonstrate the utility of networks scored against multiple Gold Standards for the pre-diction of Gene Ontology annotations for genes from KEGG pathways. Finally, we apply the networks to the functional prediction of genes which were uncharacterised in datasets from 2007, and evaluate the network results in the light of recent annotations.

About the authors

Katherine James
School of Computing Science, Newcastle University, Newcastle upon Tyne NE1 7RU, United Kingdom
Email: katherine.james@ncl.ac.uk

Samantha J. Lycett
Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom
Email: slycett@staffmail.ed.ac.uk

Anil Wipat
School of Computing Science, Newcastle University, Newcastle upon Tyne NE1 7RU, United Kingdom
Email: anil.wipat@ncl.ac.uk

Jennifer S. Hallinan
School of Computing Science, Newcastle University, Newcastle upon Tyne NE1 7RU, United Kingdom
Email: j.s.hallinan@ncl.ac.uk

Suggested keywords

DATA INTEGRATION
SYSTEMS BIOLOGY
GOLD STANDARDS
FUNCTIONAL NETWORKS

Multiple Gold Standards Address Bias in Functional Network Integration

Katherine James[1, §], Samantha J. Lycett[2, §], Anil Wipat[1] and Jennifer S. Hallinan[1,*]

[1]School of Computing Science, Newcastle University, Newcastle upon Tyne NE1 7RU, United Kingdom

[2]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom

ABSTRACT

Motivation: Network integration is a widely-used method of combining large, diverse data sets. Edge weights, representing the probability that an edge actually exists, can add greatly to the value of the networks. The edge weights are usually calculated using a Gold Standard dataset. However, all Gold Standards suffer from incomplete coverage of the genome, and from bias in the type of interactions detected by different experimental techniques. Consequently the use of a single Gold Standard tends to bias the integrated network. **Results:** We describe a novel Bayesian Data Fusion method for selecting and using multiple Gold Standards for scoring datasets prior to integration. We demonstrate the utility of networks scored against multiple Gold Standards for the prediction of Gene Ontology annotations for genes from KEGG pathways. Finally, we apply the networks to the functional prediction of genes which were uncharacterised in datasets from 2007, and evaluate the network results in the light of recent annotations.

Contact: j.s.hallinan@ncl.ac.uk

1 INTRODUCTION

Cells are complex systems of interacting parts, including genes, gene products and metabolites (Kitano, 2002). The field of Systems Biology aims to gain an understanding of these complex systems in terms of how these parts function together (Barabasi *et al.*, 2004; Ideker *et al.*, 2001). Many, diverse experimental techniques have been developed to study interactions between genes and gene products. Techniques range from detection of direct physical interactions between proteins (Pagel *et al.*, 2005) to understanding of co-expression of the genes (Edgar *et al.*, 2002). The resulting data is stored in a multiplicity of online databases; the 2010 Database Issue of Nucleic Acids research reported upon more than 1,000 such databases (Cochrane *et al.*, 2010). Diverse data sources can be combined to provide a more complete view of the functional interactions occurring in the cell and to reduce the impact of experimental noise. Data integration may also lead to enhanced understanding of functional interactions between genes or proteins, by combining multiple, weak sources of evidence for interactions present in multiple data sources (Hallinan *et al.*, 2007; Joyce *et al.*, 2006; Lee *et al.*, 2004). Integrated networks can be used in a variety of ways. They can be used to infer protein function (Deng *et al.*, 2003; Karaoz *et al.*, 2004), to detect protein complexes (Brohee *et al.*, 2008; Enright *et al.*, 2002) or to predict novel interactions (Shoemaker *et al.*, 2007; Yu *et al.*, 2006).

Data from multiple techniques can be combined naively into a network in which nodes represent genes or gene products, and edges represent any type of interaction between the nodes. However, the diverse experimental techniques used to measure functional relationships each have their own strengths, weaknesses and error rates (Hart *et al.*, 2006; Sprinzak *et al.*, 2003).Potentially, therefore, a more useful network, can be constructed by taking the quality of each dataset into account. A significant challenge when integrating diverse datasets is estimating the relative importance and quality of each dataset in a consistent manner (Jansen *et al.*, 2004).

The most commonly used method for calculating dataset quality is scoring against a Gold Standard (Myers *et al.*, 2007): a reference network containing a set of interactions believed, with high confidence, to be biologically correct (Browne *et al.*, 2009). In some cases a second, negative, set of interactions that are believed not to occur in the organism is included in the Gold Standard (Smialowski *et al.*, 2010). There are two major uses for Gold Standard data. A statistical algorithm may be used to compare each dataset to the Gold Standard prior to integration of the datasets, in an effort to estimate the "goodness" of each dataset (Lee *et al.*, 2004; Li *et al.*, 2006). Alternatively, the Gold Standard data may be used to train a machine learning algorithm to recognise true interactions in the datasets (Jaimovich *et al.*, 2006; Yellaboina *et al.*, 2007). In both cases the final network of interactions is annotated with edge weights corresponding to the confidence in that interaction being correct (Kiemer *et al.*, 2007).

The quality of the Gold Standard is vital to the accuracy of conclusions drawn from network analysis (Jansen *et al.*, 2004). Reference data is therefore commonly obtained from human expert-curated databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.*, 2000), the Munich Information Center for Protein Sequences (MIPS) (Mewes *et al.*, 1997) or the Gene Ontology (GO) (Ashburner *et al.*, 2000). Alternatively, manually-curated Gold Standards can be created for a specific application (Myers *et al.*, 2006). Gold Standard networks typically represent biologically meaningful interactions of a single type, such as shared pathway (Lee *et al.*, 2004), shared biological process (Lee *et al.*, 2007), shared function (Antonov *et al.*, 2006) or shared complex membership (Franzosa *et al.*, 2009). Negative Gold Standard datasets are commonly based on cellular location (Jansen *et al.*, 2004).

Even expert-curated databases have biases in the type of data they contain, with highly-studied proteins and processes likely to be over-represented. For instance the KEGG database is restricted to proteins present in metabolic pathways, while MIPS contains data about physical protein-protein interactions. Consequently, assessing the quality of diverse experimental datasets against a

[*]To whom correspondence should be addressed.

§These authors have contributed equally to this paper

KEGG Gold Standard biases the dataset confidence scores towards metabolic pathways and therefore biases the final integrated network in the same way (Lee *et al.*, 2009). Experimental datasets are also biased due to the experimental type and design (Huttenhower *et al.*, 2008; James *et al.*, 2009). Scoring experimental datasets against different Gold Standards can therefore produce quite different results (Myers *et al.*, 2006; Yu *et al.*, 2009). For example, a dataset may score highly against a KEGG-based Gold Standard because the experimental technique used to gather the data preferentially detects metabolic interactions. However, the same dataset could score weakly against a MIPS-Complexes Gold Standard, since that Gold Standard represents a different type of interaction, complex membership. However, both types of interaction are equally important to the organism. Given the inherent bias in individual Gold Standards, we suggest that the use of multiple Gold Standards should reduce bias in the final integrated network.

We present a novel Bayesian Data Fusion method for the construction of integrated networks. Multiple, diverse Gold Standards are selected in a principled manner, experimental interaction datasets are scored against these Gold Standards, and the datasets are integrated using our Bayesian approach. Reference data derived from KEGG, GO and MIPS were used together with experimental *Saccharomyces cerevisiae* datasets from the BioGRID database to create a composite integrated functional network. The utility of this network for the prediction of gene function was confirmed for genes on known pathways. Finally, functional predictions were made for uncharacterised genes in datasets from 2007. A comparison with current annotations demonstrated that the integrated functional network was able to predict biological functions at a higher level of detail than previously described methods.

2 METHODS

2.1 Datasets

Two types of networks were used to create the integrated network: Experimental and Gold Standard.

Version 30 of the BioGRID[1] database was used as the source of Experimental data. BioGRID contains both high-throughput and smaller, manually-curated datasets covering 22 experimental categories. During the initial evaluation the data was split by experimental type. Due to the range of dataset sizes the data were subsequently split by individual study for the final network integration. Studies containing 100 or more interactions were designated as individual datasets, while studies with fewer than 100 interactions were combined by experiment type.

Nineteen candidate Gold Standard networks were generated, from which four final Gold Standards were selected, as described below. The Gold Standard datasets were drawn from three manually curated databases, widely used as Gold Standards: KEGG, GO, and MIPS (Table 1). The candidate Gold Standard networks were created using data from each of the three sources, using the release current on the release date of BioGRID version 30.

Two networks were generated from the KEGG Pathway database[2]: a network designated *KEGG*, in which all genes in the same pathway are assumed to be functionally associated; and one called *KEGG DIRECT*, in which only genes sharing the same enzyme classification, involved in the same reaction and on the same pathway are assumed to be connected.

Fifteen candidate Gold Standard networks were generated from the three branches of the GO database[3]: Biological Process (*BP GO*), Molecular Function (*MF GO*) and Cellular Compartment (*CC GO*). GO has a hierarchical organisation, and different genes are annotated with terms at different depths within the hierarchy. We created networks in which nodes represent genes and edges represent a common GO annotation at depth *N* or lower. Gold Standard networks were created for *N* ranging from 5 to 9, for each branch of the hierarchy.

Two candidate Gold Standard networks were created from the MIPS Comprehensive Yeast Genome Database[4] using data from the `Enzyme Class` and `Complexes` catalogs. In the *Enzymes* network an edge represents a common enzyme classification, while in the *Complexes* network an edge represents membership of the same complex.

Table 1 Candidate Gold Standard Networks.

Source Data	Gold Standard	Link Types
KEGG (2 networks)	KEGG KEGG DIRECT	Genes in the same pathway Genes have same enzyme classification AND enzymes associated in same reaction AND reaction on same pathway
GO (15 networks)	BP GO *N*	Genes have common annotation to biological process at level *N* or below, *N* = 5 – 9
	CC GO *N*	Genes have common annotation to cellular component at level *N* or below, *N* = 5 – 9
	MF GO *N*	Genes have common annotation to molecular function at level *N* or below, *N* = 5 – 9
MIPS (2 networks)	MIPS Complexes	Genes in the same complex
	MIPS Enzymes	Genes have the same enzyme classification

2.2 Dataset Bias

The Gold Standard networks are derived from different types of experiment, and hence should be dissimilar. To test this assumption, we developed a log likelihood similarity measure:

$$\Lambda = \frac{L_1}{L_0} = \frac{p_{1,1}}{p_{0,1}} \times \frac{p_{0,0}}{p_{1,0}} \\ = \left(\frac{n_{1,1}}{N_1} \times \frac{N_0}{n_{1,0}} \right) \times \left(\frac{n_{0,0}}{N_0} \times \frac{N_1}{n_{0,1}} \right) = \frac{n_{1,1} n_{0,0}}{n_{1,0} n_{0,1}} \qquad (1)$$

where $n_{1,1}$ is the number of interactions in both network *D* and Gold Standard network *G*; $n_{1,0}$ is the number of interactions *D* that are not in *G*; $n_{0,1}$ is the number of interactions in *G* that are not in *D;* $n_{0,0}$ is the number of interaction not present in either *D* or *G*; N_1 is the total number of positives in *G*; L_1 is the likelihood that *D* network *D* measures ‘true’ links against the gold standard network *G*:

$$L_1 = \frac{p_{1,1}}{p_{1,0}} \quad , \qquad (2)$$

[1] http://thebiogrid.org/

[2] http://www.genome.jp/kegg/pathway.html

[3] http://www.geneontology.org/

[4] http://mips.helmholtz-muenchen.de/genre/proj/yeast/

and L_0 is the likelihood that network *D* measures ‘false’ links:

$$L_0 = \frac{p_{1,0}}{p_{0,0}} \tag{3}$$

A log-likelihood ratio score, *ln(*Λ), of 0 indicates that the networks are neither similar nor dissimilar, a large positive score means that the networks are very similar, and a large negative score means that the networks are very dissimilar.

2.3 Gold Standard Selection

Gold Standards were selected from the candidate set in order to:

(1) have low similarity to each other; and

(2) produce a range of similarity scores across the experimental data sets.

To address criterion 1, similarity scores were computed for each pair of Gold Standards using Equation 1, and candidate Gold Standards were grouped by similarity. From each group of related networks, the Gold Standard network with the largest number of interacting genes was selected, in order to optimise coverage of the data (Figure 2).

To address criterion 2, each Experimental dataset was scored against the remaining candidate Gold Standards using Equation 1. To identify the maximally different Gold Standards, a Principal Components Analysis (PCA) was applied to the dataset scores, and was repeated using all the experimental datasets but excluding each reference network in turn. Four Gold Standard networks were finally selected.

2.4 Network Integration

A three-stage data integration process was used. The Experimental datasets were first scored against each of the four Gold Standard networks. Then the scores for the individual datasets against each Gold Standard were combined, giving each edge a vector of four probabilities. The elements of the vector correspond to the probability of a functional interaction, as measured by each of the Gold Standards. Finally, the scores for each protein pair were combined into a single probability representing the confidence in the interaction.

2.4.1. Scoring Networks. Scoring was performed using Bayes’ theorem in odds ratio form:

$$O(L \mid D) = \frac{P(L \mid D)}{P(L^C \mid D)} = \frac{P(D \mid L)p(L)}{P(D \mid L^C)P(L^C)} = \frac{L_1}{L_O} O(L) \tag{4}$$

where O(L|D) is the posterior odds of links being present given that data D was measured vs. links not present (L^c) given that data D was measured; L_1 is the likelihood of links being present; L_0 is the likelihood of links not being present and O(L) is the prior odds of links being present vs. links not present.

The posterior probability of the links being present, P(L|D) can be recovered from the posterior odds:

$$O(L \mid D) = \frac{P(L \mid D)}{P(L^C \mid D)} = \frac{P(L \mid D)}{1 - P(L \parallel D)} \tag{5}$$

$$P(L \mid D) = \frac{O(L \mid D)}{1 + O(L \mid D)} \tag{6}$$

If a link between nodes *i* and *j* is measured in many conditionally independent data sets:

- Let d_{ij}^k represent the value of a link between nodes *i* & *j* in data set *k*
- Let a_{ij} represent the value of a link between nodes *i* & *j* in the integrated network
- Let D_k represent a collection of *k* data sets

If a value for a_{ij} has already been established from *k-1* data sets (D_{k-1}), and a new data set *k* provides more evidence for this interaction, we combine the new evidence with the existing evidence. The posterior odds for a link between node *i* and node *j* given the *k* data sets is thus:

$$\begin{aligned} &O(a_{ij} \mid D) = O(a_{ij} \mid d_{ij}{}^{k}, D_{k-1}) \\ &= \Lambda_k (d_{ij}{}^{k} \mid a_{ij}) O(a_{ij} \mid D_{k-1}) \end{aligned} \tag{7}$$

That is, the posterior odds for the link given the new evidence and all of the previous evidence is equal to the likelihood ratio score of the new evidence (k^{th} data set) multiplied by the odds for the link, considering the previous evidence (*k-1*th ... *1*st data sets). Decomposing the odds for the *k-1*th to *1*st data sets into the component likelihood ratios, and taking natural logs of both sides results in:

$$\ln(O(a_{ij} \mid D)) = \sum_{k=1}^{N} \ln(\Lambda_k) + \ln(O_{init}(a_{ij})) = K \tag{8}$$

The actual posterior probability (p_{ij}) for the link between nodes *i* and *j* being present given all the data and prior assumptions is then:

$$p_{ij} = \frac{\exp(K)}{1 + \exp(K)} \tag{9}$$

2.4.2 Network Integration with Priors. The prior odds is the ratio of the prior probability that a link exists between nodes *i* and *j*, divided by the prior probability that no link exists (Mukherjee *et al.*, 2008). Links present in the Gold Standard network are expected to have a high probability of actually being present in nature, since the Gold Standard is by definition a well established and manually-curated source of reference data for the biological community. High log prior odds should therefore be used for those links present in the Gold Standard. However, if a link is not present in the Gold Standard it may exist but may not have been identified. Therefore, interactions not present in the Gold Standard are not necessarily true negatives. Consequently, we set the prior odds for absent links to be 1.0, and so the log-odds are 0.0. Since the edges in the *KEGG*, *MF GO* and *MIPS Enzymes* Gold Standard networks represent known functional interactions, these edges are included in their respective integrations as prior information with a high probability.

In the case of the *CC GO* Gold Standard, it is assumed that co-located proteins are more likely to interact than those in different cellular compartments. However, it is not true that all proteins annotated to the same cellular compartment have a functional interaction. For this reason, the log prior odds for the integration with respect to Cellular Component GO were set to 0.0 for all the links.

2.4.3 Integration of the Composite Networks. The edge weights for each dataset against each of the four selected Gold Standards (section 2.3) were not strongly correlated (data not shown). We therefore assumed that the combined probabilistic networks are conditionally independent for the purposes of the final integration.

If the networks are conditionally independent, the final probability for a functional interaction between nodes *i* and *j* is the probability of that link under the *KEGG* combined network, OR the *CC GO* combined network, OR the *MF GO* combined network, OR the *MIPS Enzymes* combined network. The logical OR operation is appropriate because an inclusive definition of functional interaction was used. The final link probability under an OR operation is calculated as one minus the probability of the link not existing in any of the networks:

$$p_{Final} = 1 - (1 - p_{KEGG})(1 - p_{CCGO})(1 - p_{MFGO})(1 - p_{Enzymes}) \quad (10)$$

Thus, even if the individual probabilities are very weak, integrating weak lines of evidence still produces a higher final link probability.

2.5 Network Evaluation

The final integrated network contains probabilistic links between genes or their products. Since the links represent functional interactions, it can be assumed that the function of a particular gene is related to the functions of the neighbours to which it is joined by high probability edges: the 'guilt-by-association' principle (Oliver, 2000). To test the ability of the final integrated network to correctly predict gene function, the functions of the interaction partners of genes with known annotations were examined using the *Majority Rule* (Schwikowski *et al.*, 2000). In this algorithm, for each gene of interest:

- All the interaction partners with probability of interaction >= 0.9 are selected;
- The interaction partners are sorted by frequency of occurrence to find the most popular and most specific GO term for each branch of GO;
- The most highly-represented GO term in each GO category, or alternatively the five most popular GO terms in each category are used as the final functional predictions.

The predictions were compared with known annotations for three KEGG pathways which have been previously analysed using functional linkage networks (Myers *et al.*, 2005): SCE03010 (Ribosome), SCE04111 (Cell cycle) and SCE00193 (ATP synthesis). Finally, the four individual Gold Standard networks and the composite network were used to produce functional annotations for unannotated genes. New predictions with probability >= 0.9 were evaluated by comparison with Gene Ontology annotations from June 2010.

3 RESULTS

3.1 Similarity of Experimental datasets

A heat map image of the log-likelihood ratio score between pairs of BioGRID datasets is shown in Fig. 1. The log-likelihood ratio has been calculated over the nodes (genes) common to both networks in each pair.

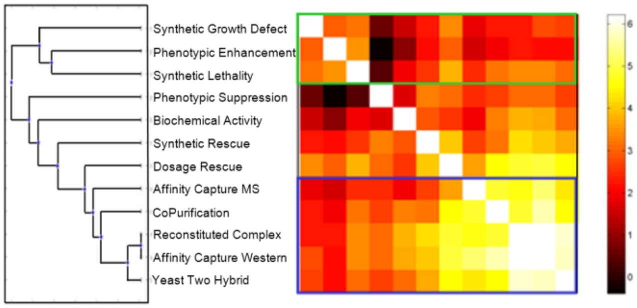

Fig. 1. Groupings of BioGRID data networks according to the log-likelihood ratio score over sub-sets of common genes.

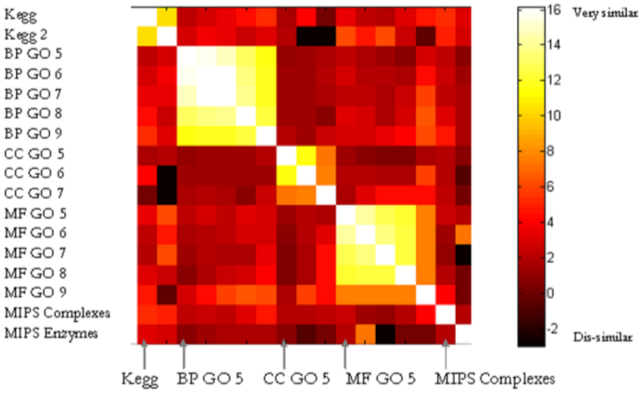

Fig. 2. Log likelihood score heat map for the reference networks.

This dendrogram confirms that networks constructed from physical interaction experiments are similar to each other and that networks from genetic interaction experiments are similar to each other, and are less similar to the physical interaction networks.

3.2 Gold Standard Selection

Gold Standards were selected from the set of candidates in a two-stage process. The similarity scores for the candidate Gold Standards are shown in Fig. 2. One Gold Standard was selected from each group of similar networks. Specifically, networks were discarded if they were too small to have sufficient overlap with the data or too large to provide a good measure of experimental false positives. This process produced a shortlist of five candidate reference networks: *KEGG*; *BP GO 6*; *CC GO 5*; *MF GO 5*; and *MIPS Enzymes*.

PCA was then applied to the dataset scores for each Gold Standard. The resulting eigenvectors are shown as a heat map in Fig. 3, with the colour of the cells representing the strength of each eigenvector in each network, ranging from low (black) to high (white). The *KEGG*, *MIPS Enzymes* and *MF GO 5* were the most important datasets in the first three eigenvectors.

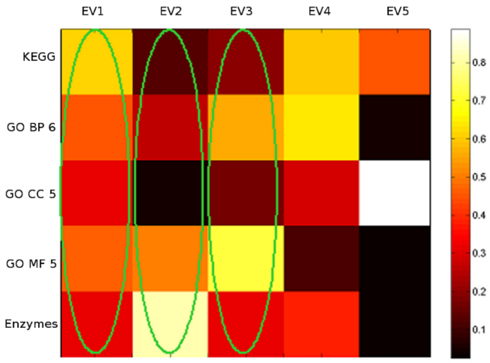

Fig. 3. Importance of the first five eigenvectors (EV1 to EV5) of the scores of the experimental datasets against the five candidate Gold Standards.

Table 2 Number of times each reference network contributes the most to each eigenvector. The final Gold Standards selected are indicated in bold.

Gold Standard	EV 1	EV 2	EV 3	EV 4
KEGG	**6**	1	1	1
BP GO 6	0	1	5	2
CC GO 5	0	0	0	**8**
MF GO 5	1	2	**6**	1
MIPS Enzymes	5	**8**	0	0

PCA was repeated using all experimental datasets and excluding each reference network in turn. Based on the consensus results (Table 2) *KEGG*, *MIPS Enzymes*, *MF GO 5* and *CC GO 5* were selected as the final Gold Standards.

3.3 Network Integration

The final integration of the Experimental datasets scored against the *KEGG*, *CC GO 5*, *MF GO 5* and *MIPS Enzymes* Gold Standards was performed using the three-stage method described in section 2.4.

3.4 Network Validation

Functional predictions were produced for genes already annotated to three KEGG pathways: SCE03010:Ribosome; SCE04111:Cell cycle; and SCE00193:ATP synthesis, using the Majority Rule with an edge probability cut-off of 0.9. For the three pathways, the predicted GO annotations for all the genes on the pathway were compared across the Biological Process, Cellular Component and Molecular Function categories, and the number of exact matches was recorded. Fig. 4a displays the percentage of genes on the pathway with three or fewer exactly correct predictions for the most popular GO term. In Fig. 4b the top five most popular terms are considered. The percentage of genes with at least one correctly predicted annotation in the top five predictions is more than 80% in all three pathways.

3.5 Predictions for Genes with Unknown Functions

We then tested the integrated network for its ability to predict the function of genes which were un-annotated at the time the data comprising the network was released, but which were annotated in subsequent data releases. All un-annotated genes involved in at least one interaction were selected from the June 2007 data and annotations were predicted for each using the Majority Rule with an edge weight cutoff of 0.9. The new annotations were evaluated by comparison with *Saccharomyces* Genome Database (SGD) annotations from June 2010 (Dwight *et al.*, 2002). Predictions to obsolete terms were replaced with their current equivalent term. This prediction process was repeated for networks scored against each of the four Gold Standards only, giving a total of five different sets of predictions. The network scored against all four Gold Standards is henceforward referred to as the composite network, while the networks scored against only one Gold Standard are named according to the Gold Standard against which they were scored.

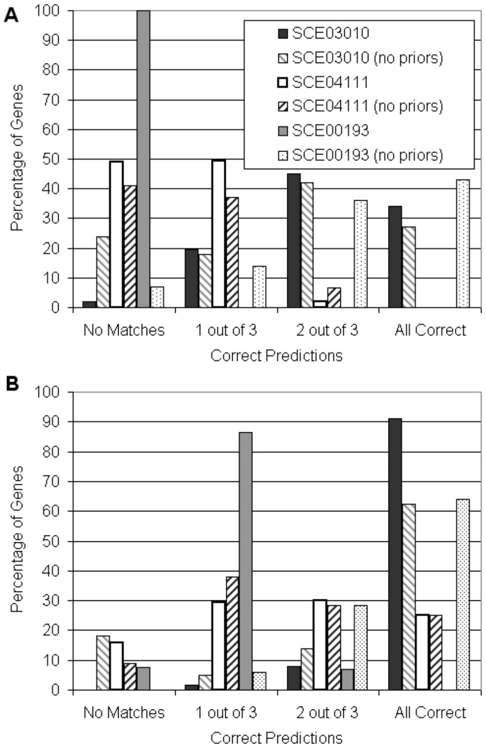

Fig. 4. Percentages of exact GO term (A) and top 5 (B) prediction matches for all the genes in three selected KEGG pathways.

Gene Ontology annotations are of two types: manually curated and computationally inferred. Curated annotations are generally considered to be of higher quality and confidence. Of the 209 un-annotated genes in the June 2007 dataset, 91 are still, as of June 2010, un-annotated in all three branches of GO. Of those which have been annotated in the interim 31 had curated annotations to at least one branch of GO. Twenty-five of which were annotated with Biological Process, sixteen with Molecular Function, and sixteen with Cellular Compartment. Only five genes had curated annotations to all three branches of GO. The number of predictions produced by the networks differed, with the composite network being the only one producing predictions for all 209 genes.

We used two criteria to assess the quality of the predictions for the 31 recently-annotated genes, based on whether a prediction was an exact match, inexact match, or non-match to the curated annotation. A match was taken to include the exact annotation or any child of that term. Predictions were indirect matches if the predicted annotation was biologically consistent with known annotations.

The percentage of genes with at least one prediction matching the known curated annotations generated by the composite network was at least double that of the single Gold Standard networks for both direct and indirect matches. In total, 71% of the curated genes had a matching prediction from the composite network.

While the majority of the genes lacked curated annotations, computational annotations were available for 118 genes. The predictions for these genes were compared with all known annotations, including those of lower confidence, using the same

Table 3 Summary of the predicted matches to known annotations. The column headed 'Genes' indicates the number of genes for which each network made predictions above the cutoff level of confidence.

		Curated (31 genes)		Computational (118 genes)	
Network	Genes	Direct No. (%)	Direct + Indirect No. (%)	Direct No. (%)	Direct + Indirect No. (%)
Composite	209	13(41.9)	22 (71.0)	69 (58.5)	80 (67.8)
MF	60	5(16.1)	10 (32.3)	26 (22.0)	35 (30.0)
CC	93	3(9.7)	6 (19.4)	17 (14.4)	27 (22.9)
MIPS	40	3(9.7)	6 (19.4)	12 (10.2)	16 (13.6)
KEGG	135	5(16.1)	11 (35.5)	33 (28.0)	44 (37.3)

direct/indirect match criteria (Table 3). The percentage of composite network prediction matches to known annotations was at least around double that of the single Gold Standard networks. In total, 67.8% of the 118 annotated genes had matching predictions produced by the composite network.

1 DISCUSSION

Experimental datasets are known to suffer from problems of incompleteness and bias. The integration of multiple datasets, generated using a variety of approaches, is one widely-used approach to overcoming these issues. Assessing the "goodness" of experimental datasets prior to integration, in terms of genome coverage and proportion of false positive and false negative interactions reported, can also add to the usefulness of the final integrated networks. However, individual Gold Standards, even the best of the manually-curated datasets, suffer from similar problems to experimental data.

We have described an approach to dataset scoring using multiple Gold Standards, chosen in a principled manner, in order to maximise their coverage of the genome and the diversity of interactions which they contain. We demonstrate that our composite networks, scored against multiple Gold Standards prior to integration, perform significantly better than integrated networks scored against a single Gold Standard on the task of inferring functional annotations for genes.

Integrated networks are used for a variety of purposes, many of which—such as cluster analysis—are largely subjective, their value dependent upon the needs and prior knowledge of the investigator. Network analysis is often performed in an interactive, exploratory manner. Assessing the goodness of a network is therefore not straightforward. However, it is clear that some datasets are of better quality than others, having greater coverage of the genome, fewer false positives, or having been generated using more reliable technology. Datasets generated using different techniques are biased towards the detection of different types of interaction. These issues also affect Gold Standard datasets, and it has been demonstrated that experimental datasets score differently depending upon the Gold Standard used (Myers et al., 2006; Yu et al., 2009). It is therefore important to generate integrated networks which are as complete and unbiased as possible.

We chose to measure the quality of our integrated networks using prediction of function, because this usage is quantifiable; the biological function of nodes with known annotation can be predicted from their neighbours, and the existing annotation compared with the prediction. Further, functional predictions can be made for genes without annotation in one version of a dataset, and compared with annotations present in later versions of the same dataset. For evaluation we compared GO annotations from 2007 with those from 2010. Interestingly, only 118 of the genes which were not annotated in 2007 have acquired annotations in the subsequent three years.

Computationally-predicted annotations are generally assumed to be less reliable than manually curated annotations (Friedberg, 2006). However, our integrated networks were in most cases slightly more successful at predicting the computational annotations than the manual ones. It is possible that computational function prediction methods tend to predict the same sorts of functions, biased towards the data on which they were trained. However, GO uses a variety of different algorithms for these predictions all of which are different from the one used here. Another confounding factor is the fact that many proteins have multiple functions. As of August 8, 2010, SGD contains 4,038 gene products annotated to 14,853 GO terms, an average of 3.7 annotations per gene product. Since the algorithm we used annotates a protein to a single biological function, information is inevitably lost.

The individual Gold Standards which were selected by our approach as being most complete and diverse include, unsurprisingly, those most widely used by the data integration community. Notable, the composite network generates high confidence predictions for all 209 genes, whereas the networks scored against single Gold Standards only produce high confidence predictions for a subset of the genes. The subsets predicted confidently by each network overlap, but not extensively. It would appear that, as hypothesized, different biases in different Gold Standards affect the performance of the networks scored against them.

Second in power to our composite network was that scored against KEGG, an intensively manually-curated dataset. The composite network made exact predictions for 13 out of 34 manually-annotated genes, compared with KEGG (and GO Molecular Function) with exact predictions for 5 out of 31. Interestingly, the composite network was exactly correct for 69 out of 118 computationally-inferred annotations, whereas KEGG was correct for only 33 genes.

The third most accurate predictor was a network scored against the MIPS dataset. This dataset is also manually curated, although not to the same extent as KEGG. The worst performer was a dataset scored against the GO Cellular Compartment database. Information about the physical co-localisation of proteins is likely to be valuable insofar as proteins which are not in the same compartment are unlikely to interact; however, the value of the database is limited, since proteins which are co-located will not necessarily interact. Also, many genes are dispatched to multiple cellular locations.

2 CONCLUSIONS

Gold Standards for assessing the value of experimental datasets suffer from a number of practical problems. One of the major issues is bias: different experimental techniques will preferentially detect different types of interactions. We describe an approach addressing this problem using multiple Gold Standards, selected to cover as much of the genetic diversity of the organism of interest as possible. We compared our approach to networks scored against a single Gold Standard by using the Majority Rule algorithm to infer GO Biological Function annotations for proteins in the network.

Our composite network, scored against four different Gold Standards, performed best at this task. The accuracy of predictions made by networks scored against single Gold Standards reflected the degree of manual curation of the datasets, and therefore supports the assumption of many researchers in this area, that the more highly manually curated a dataset is, the more reliable it is likely to be.

It would appear that the use of multiple Gold Standards, carefully chosen to provide as much coverage and diversity as possible does indeed overcome some of the problems of incompleteness and bias which plague individual Gold Standards.

ACKNOWLEDGEMENTS

We thank Malcolm Farrow, Matthew Pocock, Phil Lord and the Newcastle University Bioinformatics Support Unit for their helpful suggestions and advice.

Funding: This work was supported by the Biotechnology and Biological Sciences Research Council [BBF0060631 to AW] and the Research Councils of the UK [Academic Fellowship to JSH].

Conflicts of Interest: none declared.

REFERENCES

Antonov, A.V. *et al.* (2006) A systematic approach to infer biological relevance and biases of gene network structures, *Nucleic Acids Res*, **34**.

Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat Genet*, **25**, 25-29.

Barabasi, A.-L. *et al.* (2004) Network biology: understanding the cell's functional organization, *Nat Rev Genet*, **5**, 101-113.

Brohee, S. *et al.* (2008) NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways, *Nucleic Acids Res*, **36**, 444-451.

Browne, F. *et al.* (2009) GRIP: A web-based system for constructing Gold Standard datasets for protein-protein interaction prediction, *Source Code Biol Med*, **4**, 2-2.

Cochrane, G.R. *et al.* (2010) The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources, *Nucleic Acids Res*, **38**, 1-4.

Deng, M. *et al.* (2003) An integrated probabilistic model for functional prediction of proteins. *Proceedings of the Annual International Conference on Computational Molecular Biology, RECOMB*. 95-103.

Dwight, S.S. *et al.* (2002) *Saccharomyces* Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO), *Nucleic Acids Res*, **30**, 69-72.

Edgar, R. *et al.* (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, *Nucleic Acids Res*, **30**, 207-210.

Enright, A.J. *et al.* (2002) An efficient algorithm for large-scale detection of protein families, *Nucleic Acids Res*, **30**, 1575-1584.

Franzosa, E. *et al.* (2009) Computational reconstruction of protein-protein interaction networks: algorithms and issues, *Methods Mol Biol*, **541**, 89-8100.

Hallinan, J.S. *et al.* (2007) Motifs and modules in fractured functional yeast networks, *Computational Intelligence and Bioinformatics and Computational Biology, 2007. . IEEE Symposium on*, 189-196.

Hart, G.T. *et al.* (2006) How complete are current yeast and human protein-interaction networks?, *Genome Biol*, **7**, 120-120.

Huttenhower, C. *et al.* (2008) Assessing the functional structure of genomic data, *Bioinformatics*, **24**, 330-338.

Ideker, T. *et al.* (2001) A new approach to decoding life: systems biology, *Annu Rev Genomics Hum Genet*, **2**, 343-372.

Jaimovich, A. *et al.* (2006) Towards an integrated protein-protein interaction network: a relational Markov network approach, *J Comput Biol*, **13**, 145-164.

James, K. *et al.* (2009) Integration of full-coverage probabilistic functional networks with relevance to specific biological processes. In, *Data Integration in the Life Sciences*. 31-46.

Jansen, R. *et al.* (2004) Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction, *Curr Opin Microbiol*, **7**, 535-545.

Joyce, A.R. *et al.* (2006) The model organism as a system: integrating 'omics' data sets, *Nat Rev Mol Cell Biol*, **7**, 198-210.

Kanehisa, M. *et al.* (2000) KEGG: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Res*, **28**, 27-30.

Karaoz, U. *et al.* (2004) Whole-genome annotation by using evidence integration in functional-linkage networks, *Proc Natl Acad Sci U S A*, **101**, 2888-2893.

Kiemer, L. *et al.* (2007) WI-PHI: a weighted yeast interactome enriched for direct physical interactions, *Proteomics*, **7**, 932-943.

Kitano, H. (2002) Computational systems biology, *Nature*, **420**, 206-210.

Lee, I. *et al.* (2004) A probabilistic functional network of yeast genes, *Science*, **306**, 1555-1558.

Lee, I. *et al.* (2007) An improved, bias-reduced probabilistic functional gene network of baker's yeast, , *PLoS ONE*, **2**.

Lee, I. *et al.* (2009) Effects of functional bias on supervised learning of a gene network model, *Methods Mol Biol*, **541**, 463-475.

Li, J. *et al.* (2006) A framework of integrating gene relations from heterogeneous data sources: an experiment on *Arabidopsis thaliana*, *Bioinformatics*, **22**, 2037-2043.

Mewes, H.W. *et al.* (1997) MIPS: a database for protein sequences, homology data and yeast genome information, *Nucleic Acids Res*, **25**, 28-30.

Mukherjee, S. *et al.* (2008) Network inference using informative priors, *Proc Natl Acad Sci U S A*, **105**, 14313-14318.

Myers, C.L. *et al.* (2006) Finding function: evaluation methods for functional genomic data, *BMC Genomics*, **7**, 187-187.

Myers, C.L. *et al.* (2005) Discovery of biological networks from diverse functional genomic data, *Genome Biol*, **6**, R114.

Myers, C.L. *et al.* (2007) Context-sensitive data integration and prediction of biological networks, *Bioinformatics*, **23**, 2322-2330.

Oliver, S. (2000) Guilt-by-association goes global, *Nature*, **403**, 601-603.

Pagel, P. *et al.* (2005) The MIPS mammalian protein-protein interaction database, *Bioinformatics*, **21**, 832-834.

Schwikowski, B. *et al.* (2000) A network of protein-protein interactions in yeast, *Nat Biotechnol*, **18**, 1257-1261.

Shoemaker, B.A. *et al.* (2007) Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners, *PLoS Comput Biol*, **3**.

Smialowski, P. *et al.* (2010) The Negatome database: a reference set of non-interacting protein pairs, *Nucleic Acids Res*, **38**, 540-544.

Sprinzak, E. *et al.* (2003) How reliable are experimental protein-protein interaction data?, *J Mol Biol*, **327**, 919-923.

Yellaboina, S. *et al.* (2007) Inferring genome-wide functional linkages in *E. coli* by combining improved genome context methods: comparison with high-throughput experimental data, *Genome Res*, **17**, 527-535.

Yu, J. *et al.* (2009) Combining multiple positive training sets to generate confidence scores for protein-protein interactions, *Bioinformatics*, **25**, 105-111.

Yu, J. *et al.* (2006) Computational approaches for predicting protein-protein interactions: a survey, *J Med Syst*, **30**, 39-44.