SUPPLEMENTAL MATERIAL FOR:
Implications of Functional Similarity for Gene Regulatory Interactions

Kimberly Glass[1,2,*], Edward Ott[2], Wolfgang Losert[2,3], Michelle Girvan[2,3]

1 Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA
2 Physics Department, University of Maryland, College Park, MD, USA
3 Institute for Physical Science and Technology, University of Maryland, College Park, MD, USA

∗ E-mail: kglass@jimmy.harvard.edu

1 Attributes of the Gene Ontology and Other Functional Databases

Many other functional classification schemes have been proposed besides the Gene Ontology [20][19][18][3][9]. Here we choose two to serve as a contrasting comparison to the Gene Ontology database: (1) the *E. coli* genome and proteome (GenProtEC) database [18], and (2) the Clusters of Orthologous Groups of proteins (COG) database [20]. We chose GenProtEC since we are focusing our investigation on *E. coli* and this database is specifically dedicated to the functions performed by this organism. We chose COG as an example of a database which records gene properties in a manner which should be, on the whole, distinctly different from the Gene Ontology.

We downloaded annotation information from these databases' corresponding websites [1][2] and used it to construct two bipartite graphs, one for each database, in the same manner as with *E. coli* annotations in the Gene Ontology (see main text, Section 1.2.1). These two databases are, in general, smaller than the Gene Ontology. GenProtEC contains 23137 annotations from 3361 genes to 594 functional categories (a density of just over 1%). COG assigns each gene to one orthologous group (although some orthologous groups contain several genes). This results in 3450 annotations from 3450 genes to 2131 orthologous groups (a density of only 0.05%). This is in contrast to the Gene Ontology with a total of 119936 gene-term annotations between 3794 *E. coli* genes and 3882 functional categories (a density of 0.8%).

Figure S1 shows the degree distribution for "terms" and genes in the Gene Ontology as well as these two databases. In all three, the degree distribution of the functional categories (or orthologous groups in COG) is heavy-tailed. This reinforces our belief that taking into account the degree of a functional category is important when designing a measure to accurate reflect the functional similarity between two genes. COG's construction implies that every gene has the same degree, however, is it interesting that the degree distribution of genes both in the Gene Ontology and GenProtEC have the same basic behavior.

We used the bipartite graphs we constructed from these two databases to calculate a scaled similarity and Kappa statistic between pairs of genes. The smaller database size and sparse construction of COG are evident in the results. Using annotations from the GenProtEC database we calculated a scaled similarity and Kappa statistic for 3086481 out of a possible 5646480 gene-pairs (55%) but using annotations from COG we could only calculate a scaled similarity and Kappa statistic for 2848 of a possible 5949525 gene-pairs (0.048%). This is, again, in contrast to the Gene Ontology where we can estimate a scaled similarity and Kappa statistic for 6713626 of a possible 7195321 gene-pairs (93%). By default, any gene-pair without a score is given a default value of zero.

We calculated the maximum F-score for both the scaled similarity and Kappa statistic in each of these databases using RegulonDB as our gold standard (see the main text, Section 2.2, for more information on how we calculated the F-Score). Because of the different percentage of edges each database can assign a score, the absolute value of the F-score varies quite broadly. For example, using annotations from COG we can estimate a functional score for fewer edges than actually appear in our gold standard (and only a subset of these extend from a transcription factor). As a result, the maximum F-score in this database occurs when only a very few edges (18 and 147 for the scaled similarity and Kappa statistic, respectively), are used to define the "true positive" and "false negative" classes.

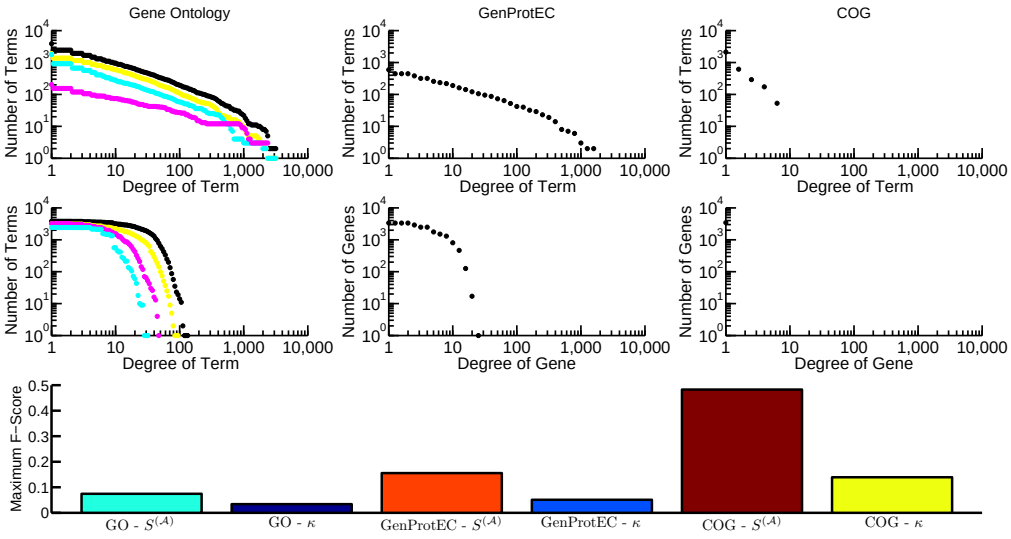


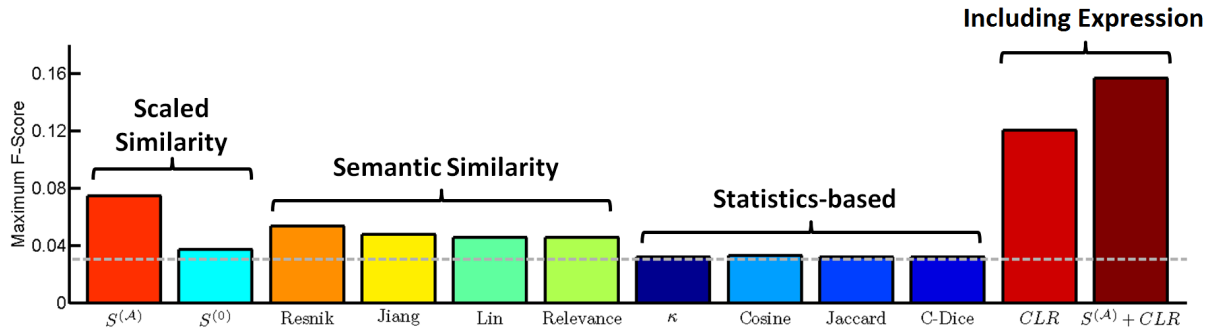
Figure S1: The annotation properties of other databases as well as the predictive power of the scaled similarity when calculated using annotations from these other databases. As a comparison the predictive power of the Kappa Statistic is also shown.

Even though the F-score between databases varies quite widely, the relative performance of the scaled similarity to the Kappa statistic is consistent and the F-score is always $2-4$ times higher for the scaled similarity compared to the Kappa statistic. This shows that, even when using other databases, adequately accounting for the degree distribution of the assigned classes is vital when constructing a similarity measure that is predictive of true regulatory interactions.

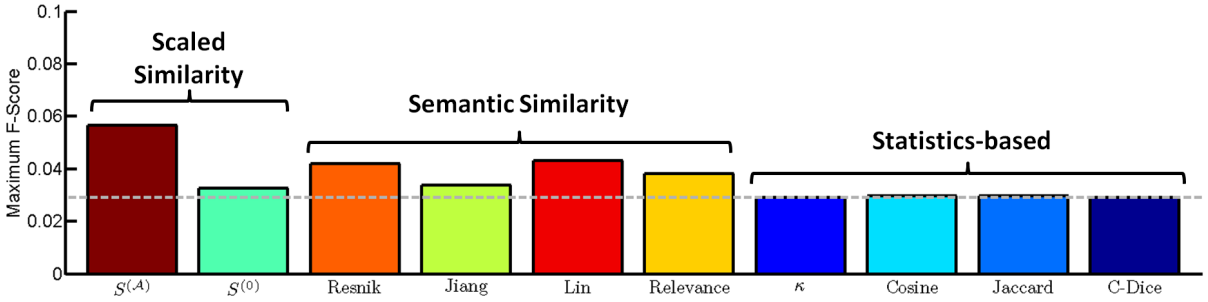
2 The Predictive Power of Functional Measures

Throughout the years many functional measures have been proposed. For a discussion of these measures see Section 1.2.2 in the main text. In Figure S2 we provide an analysis of the performance of each of these measures in regards to their ability to correctly predict "true" regulatory interactions. The predictive power was calculated by determining the maximum F-Score for each measure based on a "gold-standard" (see the main text, Section 2.2, for more information on how we calculated the F-Score). In *E. coli* we used edges reported by RegulonDB [6] as our "gold-standard" and for yeast, edges defined by ChIP-chip [12] (see the main test, Section 3.5 for more information on how we used the ChIP data to create a gold standard). The color of each bar corresponds to how predictive each measure is relative to the other measures, with dark red representing the strongest predictive power and dark blue the weakest. The measures presented in this figure fall into several classes. $S^{(\mathcal{A})}$ and $S^{(0)}$ are calculated based on the *scaled similarity* measure developed in this paper (see the main text, Section 2.1-2.2 for more information on how we calculated the scaled similarity and determined $\mathcal{A}$). *Semantic similarity* measures were first applied to the Gene Ontology by Lord et. al. in 2003 [11]. Semantic similarity measures utilize the GO hierarchy to determine the similarity between terms and then genes. Four separate measures which utilize this approach are presented here: the one based on Resnik [16], Lin [10], Jiang and Conrath [8], and finally the Relevance measure developed by Schlicker et. al. which combines the methods of Resnik

and Lin [17]. *Statistics-based* approaches calculate a functional similarity between two genes largely based on the number of shared annotations. The measures presented here include measures based on the Kappa-statistic [7], the cosine similarity [4], a weighted Jaccard index [15], and Czekanowski-index/Dice-coefficient [13]. All semantic similarity and statistics-based similarity measures were calculated using the csbl.go package in R [14]. Finally, reconstructions *including expression* data utilize microarray data to predict a gene regulatory network. We show the predictive power of the CLR network reconstruction algorithm [5], as well as the predictive power when predictions from CLR are combined with those from the scaled similarity measure (see the main text, Section 3.6, for more information on how we combined the scaled similarity with CLR).



(a) Predictive Power of Functional Measures in *E. coli*



(b) Predictive Power of Functional Measures in Yeast

Figure S2: The predictive power of many different established functional similarity measures as well as those developed in this paper. Results are shown both for (a) *E. coli* and (b) Yeast.

In our analysis we have used the maximum F-score to define a cut-off above which we believe our edges are most likely to represent true regulatory interactions. However, in the absence of a gold standard the F-score cannot be calculated and one can no longer determine a cut-off in this manner. In RegulonDB, the are 6725 edges which have the potential to be predicted by these ontology-based functional measures. Therefore, to address this issue we compared the value of the maximum F-score to the value the F-score has when using a division of the top 6725 edges to define "true positives" and "false negatives". As Figure S3 demonstrates, the conclusions that can be made based off this set cut-off vary little from those made by taking a cutoff where the F-score is maximized. Interestingly, the "semantic similarity" metrics show the biggest decrease in predictive power. $F(N = 6725)$ for these metrics are barely distinguishable from those of the "statistics-based" metrics, which, as pointed out previously, are barely higher than random. This is largely due to the fact that the maximum F-score for the semantic similarity measures occurs at a cutoff which includes many more edges (on the order of twenty to ninety thousand) than what occurs in

regulatory networks. In contrast, more predictive measures such as the scaled similarity and CLR reach their maximum F-score very quickly, within the first few thousand edges, indicating that the top edges predicted by these methods are highly informative.

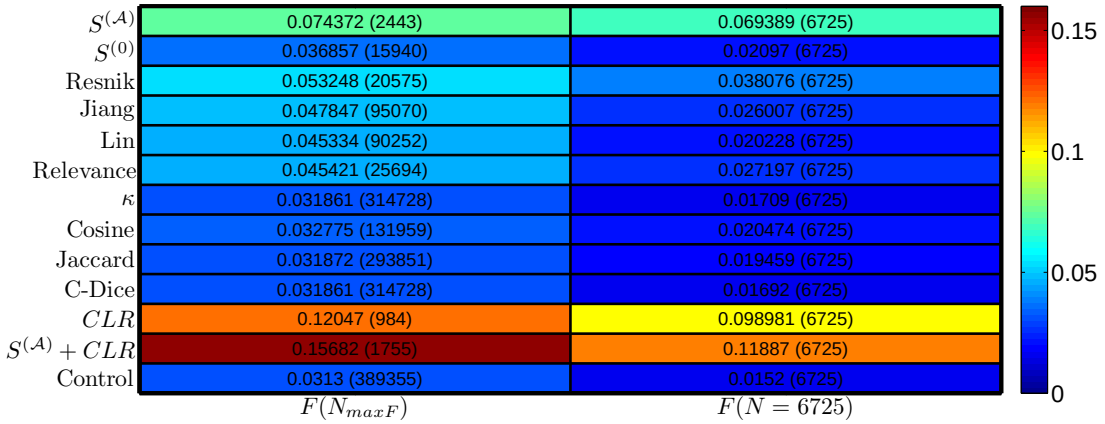


Figure S3: Comparison of the maximum F-score vs. the F-score value using a set cutoff.

References

[1] http://genprotec.mbl.edu/.

[2] http://www.ncbi.nlm.nih.gov/cog/.

[3] Ron Caspi, Tomer Altman, Joseph M. Dale, Kate Dreher, Carol A. Fulcher, Fred Gilham, Pallavi Kaipa, Athikkattuvalasu S. Karthikeyan, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Lukas A. Mueller, Suzanne Paley, Liviu Popescu, Anuradha Pujar, Alexander G. Shearer, Peifen Zhang, and Peter D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic acids research*, 38(Database issue):D473–D479, January 2010.

[4] Julie Chabalier, Jean Mosser, and Anita Burgun. A transversal approach to predict gene product networks from ontology-based similarity. *BMC Bioinformatics*, 8(1):235+, July 2007.

[5] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol*, 5(1):e8+, 2007.

[6] Socorro Gama-Castro, Veronica Jimenez-Jacinto, Martin Peralta-Gil, Alberto Santos-Zavaleta, Monica I. Penaloza-Spinola, Bruno Contreras-Moreira, Juan Segura-Salazar, Luis Muniz-Rascado, Irma Martinez-Flores, Heladia Salgado, Cesar Bonavides-Martinez, Cei Abreu-Goodger, Carlos Rodriguez-Penagos, Juan Miranda-Rios, Enrique Morett, Enrique Merino, Araceli M. Huerta, Luis Trevino-Quintanilla, and Julio Collado-Vides. Regulondb (version 6.0): gene regulation model of escherichia coli k-12 beyond transcription, active (experimental) annotated promoters and textpresso navigation. *Nucl. Acids Res.*, 36(suppl_1):D120–124, January 2008.

[7] Da W. Huang, Brad T. Sherman, Qina Tan, Jack R. Collins, Gregory W. Alvord, Jean Roayaei, Robert Stephens, Michael W. Baseler, Clifford H. Lane, and Richard A. Lempicki. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome biology*, 8(9):R183+, September 2007.

[8] J. J. Jiang and D. W. Conrath. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *International Conference Research on Computational Linguistics (ROCLING X)*, pages 9008+, September 1997.

[9] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, January 2000.

[10] Dekang Lin. An Information-Theoretic Definition of Similarity. In *In Proceedings of the 15th International Conference on Machine Learning*, pages 296–304, 1998.

[11] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble. Semantic similarity measures as tools for exploring the gene ontology. *Pac Symp Biocomput*, pages 601–612, 2003.

[12] Kenzie D. MacIsaac, Ting Wang, D. Benjamin Gordon, David K. Gifford, Gary D. Stormo, and Ernest Fraenkel. An improved map of conserved regulatory sites for saccharomyces cerevisiae. *BMC bioinformatics*, 7(1):113+, March 2006.

[13] David Martin, Christine Brun, Elisabeth Remy, Pierre Mouren, Denis Thieffry, and Bernard Jacq. GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biol*, 5(12), 2004.

[14] Kristian Ovaska, Marko Laakso, and Sampsa Hautaniemi. Fast Gene Ontology based clustering for microarray experiments. *BioData Mining*, 1(1):11+, November 2008.

[15] Catia Pesquita, Daniel Faria, Hugo Bastos, Antonio Ferreira, Andre Falcao, and Francisco Couto. Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, 9(Suppl 5):S4+, 2008.

[16] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, 1995.

[17] Andreas Schlicker, Francisco Domingues, Jorg Rahnenfuhrer, and Thomas Lengauer. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, 7(1):302+, June 2006.

[18] M. H. Serres, S. Goswami, and M. Riley. Genprotec: an updated and improved analysis of functions of escherichia coli k-12 proteins. *Nucleic Acids Research*, 32(Database issue):D300–2, 2004.

[19] M. H. Serres and M. Riley. MultiFun, a multifunctional classification scheme for escherichia coli k-12 gene products. *Microb Comp Genomics*, 5(4):205–222, 2000.

[20] Roman L. Tatusov, Natalie D. Fedorova, John D. Jackson, Aviva R. Jacobs, Boris Kiryutin, Eugene V. Koonin, Dmitri M. Krylov, Raja Mazumder, Sergei L. Mekhedov, Anastasia N. Nikolskaya, B. Sridhar Rao, Sergei Smirnov, Alexander V. Sverdlov, Sona Vasudevan, Yuri I. Wolf, Jodie J. Yin, and Darren A. Natale. The COG database: an updated version includes eukaryotes. *BMC bioinformatics*, 4(1):41+, September 2003.