We suggest that state policy makers begin by eliminating those activities that are clearly best left to federal agencies—drug approval and drug safety, for example, or situations in which national health and security are at stake, such as pandemic preparedness. Next, it's important to recognize that federal/state partnerships make considerable sense in a wide range of areas, including workforce development and research infrastructure. Academic R&D centers, for example, are often built using a combination of federal, state and private funds, a system that encourages harmonization of local, regional and federal priorities.

State-sponsored life sciences initiatives, whether related to stem cells or other, less politically charged areas of research, must be carefully crafted to avoid the pitfall of attempting to do too much with too little.

The most compelling cases for 'going it alone' with state-based policies and initiatives are centered on specific regional strengths. States with strong existing clusters in bioscience, such as California, Massachusetts and Pennsylvania, have all evolved life science initiatives that build on the strength of the existing R&D and commercial bases. California has developed interdisciplinary programs at the interface of materials, information and biosciences (QB3 initiative). Massachusetts has a $20 million Research Center Matching Fund that provides funds to build research capacity. Pennsylvania invests funds from the Master Tobacco Settlement into basic research, infrastructure and three Life Sciences Greenhouses that specialize in commercialization. That is, states that already have in place many of the essential elements—R&D capacity, ongoing commercial activity, a skilled workforce and support infrastructure—and need only fill readily identified gaps are most likely to meet with success.

State-sponsored life sciences initiatives, whether related to stem cells or other, less politically charged areas of research, must be carefully crafted to avoid the pitfall of attempting to do too much with too little. Most states are constitutionally bound to balance their budgets, and thus fiscally and physically incapable of keeping pace with scientific advancement in all areas of life sciences research. Furthermore, it would be unwise to abandon promising areas of research in favor of only politically or commercially palatable ones. Strong and coherent life sciences policies are needed to maintain a firm balance between state and federal agendas, economic development and health needs and to maximize the investments made by all in the critical area of life sciences.

Melvin L Billingsley[1,2] & Michele Washko[1]

[1]Life Sciences Greenhouse of Central Pennsylvania, 225 Market St., Suite 500, Harrisburg, Pennsylvania 17101, USA. [2]Department of Pharmacology, Penn State University College of Medicine, 500 University Dr., Hershey, Pennsylvania 17033, USA.
e-mail: mlb@lsgpa.com

1. Battelle Technology Partnership Practice. *Growing the Nation's Bioscience Sector: State Bioscience Initiatives 2006* (Battelle Technology Partnership Practice, Columbus, OH, 2006). http://www.bio.org/local/battelle2006

The art of gene function prediction

To the editor:

Determining the functions of genes and proteins is a central problem in biology, fundamental to understanding the molecular and biochemical processes that sustain health or cause disease, to identifying and validating new drug targets and to developing reliable diagnostics. Recent advances in genomic sequencing have generated

an astounding number of new putative genes and hypothetical proteins whose biological function remains a mystery. On average, as many as 70% of the genes in a genome have poorly known or no known functions[1].

Many computational formalisms have emerged for integrating different data sources for predicting protein function[2,3]. Although specific technical criticisms can be made of each of these approaches, a general problem with these formalisms is that defining them is still largely an art. In this correspondence, we respond to a specific method proposed by Vazquez *et al.* (*Nat. Biotechnol.* **21**, 697–700, 2003). Using well-established ideas in graph theory, we present an algorithm that is guaranteed to produce a maximally globally consistent functional assignment to proteins in a protein-protein interaction network. Moreover, because the number of edges in a protein interaction network is usually linear in the number of proteins, our solution runs in time quadratic in the size of the network. In practice, this algorithm computes an optimal solution within seconds for widely available protein interaction networks, such as those of *Drosophila melanogaster*, *Caenorhabditis elegans* and human. In contrast, the computational solution proposed by Vazquez *et al.* is not guaranteed to provide an optimal solution in time sub-exponential in the number of nodes in the network. Our novel algorithmic formulation allows us for the first time to compare the solutions produced by the globally optimal method with the predictions based on a simple 'guilt by association' principle. Our study suggests that the optimal solutions produced by the global methods do not improve significantly on the simpler local approaches.

A general formalization of the problem of predicting protein function uses a functional linkage network (FLN), in which each node is a protein and there is an edge between two nodes if there is evidence that the nodes may share the same function. Typical sources of these edges are protein-protein interactions, correlations in gene expression profiles, literature mining and other experimental or computational techniques. A popular approach for predicting function based on FLNs uses a simple local threshold rule (often referred to as 'guilt-by-association')[4]. This rule is based on the hypothesis that if at least some prespecified fraction of the neighbors of a given protein '*p*' in the FLN are annotated with a particular function, we might 'transfer' this functional annotation to *p*. We refer to this approach and its probabilistic variants as local consistency.

Other methods attempt to achieve a globally consistent annotation by minimizing the number of locally inconsistent functional assignments (Vazquez *et al.* and ref. 3) or by maximizing the probability of the functional assignments, given all the probabilistic constraints in the network[5]. In particular, Vazquez *et al.* formulate functional annotation as a global optimization problem

Table 1 Comparison of precision and recall

Algorithm	Precision	Recall
Local (guilt by association)	76.2%	77.5%
Global (minimum cut)	77.3%	75.2%

where they seek to assign functional labels to the nodes of the graph in such a way that the number of inconsistent edges (an edge is inconsistent if it is incident on nodes with different functional assignments) is minimized. The formalism associates a state variable 'x_i' with each node in the FLN. Each variable can take the value 1 (the corresponding protein is assigned the function under consideration) or –1 (the protein is not assigned the function under consideration). An edge between two nodes 'i' and 'j' has weight 'w_{ij}'. To minimize the number of inconsistent assignments, they minimize the energy ('E') of a system expressed by the equation below:

$$E = -\sum_i \sum_j w_{ij} x_i x_j$$

The approach proposed by Vazquez *et al.* uses simulated annealing to minimize the number of inconsistent edges. In many FLNs, including the one studied by Vazquez *et al.*, all edges have positive weights. In this situation, we observe that the minimization of global inconsistency above is equivalent to the problem of finding a minimum cut in an appropriately defined graph. In particular, we can optimize E by partitioning the nodes of the graph into two sets: a set of nodes with state 1, and a set of nodes with state –1, such that the weighted sum of edges connecting the nodes in the two sets is minimized. This problem can be solved optimally in $O(nm \log n)$ time using a min-cut/max-flow algorithm[6], where 'n' is the number of nodes in the network and 'm' is the number of edges in the network (see **Supplementary Methods** online for the derivation of the graph theoretic transformation). While this work was in review, Nabieva *et al.*[7] reported a similar application of graph cuts to function prediction using Munich Information Center for Protein Sequences (now the Institute for Bioinformatics, Munich) functional annotations.

We compared this new formulation with the commonly used guilt-by-association method and constructed an FLN from the protein-protein interactions in budding yeast (*Saccharomyces cerevisiae*) in the General Repository for Interaction Datasets (GRID)[8]. We focused our attention on 82 functions in the Gene Ontology[9] that yielded high precision and recall in our previous study[3]; proteins predicted to have these functions are good candidates for experimental validation. We tested the performance of the two methods using leave-one-out cross validation (as described in **Supplementary Methods**). To our surprise, our results suggest that the global optimization does not provide a substantial advantage over the simple guilt-by-association rule, subject to a number of qualifications described below. **Table 1** summarizes the precision and recall of the two methods averaged over all the functions we studied.

It is possible that approaches for functional annotation that either annotate FLN edges with experimentally derived measures of confidence[10] or integrate diverse multimodal sources of experimental evidence[11–13] might prove that sophisticated technical approaches lead to substantial improvement in accuracy of prediction. In the meantime, our study suggests a critical need for developing commonly accepted annotation benchmarks and evaluation methodologies for the growing number of functional prediction systems. These methodologies might be modeled after competitions, such as Critical Assessment of Techniques for Protein Structure Prediction (CASP), that evaluate protein structure prediction systems. It is also necessary that function prediction engines and the predictions themselves be made available to the research community[14]. Perhaps more importantly, our results underscore the need for community-wide experimental initiatives for validating computationally predicted functional annotations, as proposed by Roberts[15, 16]. Such efforts will enable a more direct exchange of functional predictions between experimental and computational scientists that will drive the next generation of predictions and experimental validations.

Note: Supplementary information is available on the Nature Biotechnology website.

AUTHOR CONTRIBUTIONS

T.M.M., C.-J.W. and S.K. contributed equally to this work.

T.M. Murali[1], Chang-Jiun Wu[2] & Simon Kasif[2–4]

[1]Virginia Polytechnic Institute and State University, Department of Computer Science, 660 McBryde Hall, Blacksburg, Virginia 24061, USA. [2]Boston University, Bioinformatics Program, 24 Cummington St., Boston, Massachusetts 02215, USA. [3]Boston University, Department of Biomedical Engineering, 44 Cummington Street, Boston, Massachusetts 02215, USA. [4]Children's Hospital, Boston, 300 Longwood Avenue, Boston, Massachusetts 02215, USA.
email: kasif@bu.edu

1. Enright, A.J., Kunin, V. & Ouzounis, C.A. *Nucleic Acids Res.* **31**, 4632–4638 (2003).
2. Lee, I., Date, S.V., Adai, A.T. & Marcotte, E.M. *Science* **306**, 1555–1558 (2004).
3. Karaoz, U. *et al.*. *Proc. Natl. Acad. Sci. USA* **101**, 2888–2893 (2004).
4. Schwikowski, B., Uetz, P. & Fields, S. *Nat. Biotechnol.* **18**, 1257–1261 (2000).
5. Letovsky, S. & Kasif, S. *Bioinformatics* **19**, Suppl 1, I197–I204 (2003).
6. Goldberg, A.V. & Tarjan, R.E. *J. ACM* **35**, 921–940 (1988).
7. Nabieva, E., Jim, K., Agarwal, A., Chazelle, B. & Singh, M. *Bioinformatics* **21**, Suppl 1, i302–i310 (2005).
8. Stark, C. & Tyers, M. *Genome Biol.* **4**, R23 (2003).
9. Ashburner, M. *et al. Nat. Genet.* **25**, 25–29 (2000).
10. Bader, J.S., Chaudhuri, A., Rothberg, J.M. & Chant, J. *Nat. Biotechnol.* **22**, 78–85 (2004).
11. Butte, A.J., Tamayo, P., Slonim, D., Golub, T.R. & Kohane, I.S. *Proc. Natl. Acad. Sci. USA.* **97**, 12182–12186 (2000).
12. Raychaudhuri, S., Chang, J.T., Imam, F. & Altman, R.B. *Nucleic Acids Res.* **31**, 4553–4560 (2003).
13. Jansen, R. *et al. Science* **302**, 449–453 (2003).
14. Massjouni, N., Rivera, C.G. & Murali, T.M. *Nucleic Acids Res.* **34**, W340–W344 (2006).
15. Roberts, R.J. *PLoS Biol.* **2**, E42 (2004).
16. Roberts, R.J., Karp, P., Kasif, S., Linn, S. & Buckley, M.R. *An Experimental Approach to Genome Annotation. Report* (The American Academy of Microbiology, Washington, DC, 2004).

Vazquez *et al.* respond:

The correlation between protein-protein interactions and protein function can be exploited to make protein function predictions of unclassified proteins. Kasif and colleagues revisit the protein function prediction method proposed in our paper, which exploits this correlation to provide functional annotations to unclassified proteins. Kasif and colleagues claim that methods based on global consistency do not perform better than local approaches based on the guilt-by-association method.

The method investigated by Kasif and colleagues, however, differs from the one studied in our paper. Their method analyzes one function at a time, assigning or unassigning unclassified proteins to the particular function under consideration. In contrast, the method developed by us considers all functions simultaneously, assigning unclassified proteins to different functional classes.

Kasif and colleagues use 'leave-one-out' cross-validation to quantify whether the performance of their global method is better than a local method. They conclude that

both methods have a similar performance. In contrast, the cross-validation used in our paper was the 'leave-a percent-out', where a percent of the proteins with functional annotations is assumed unclassified. In this way, we conclude that the globally consistent method results in a noticeable performance increase over a local method based on a majority rule.

In summary, the authors propose a different functional assignment method and use a different cross-validation method than those used in our paper. We conclude, therefore, that further analysis of the performance of global versus local method is warranted before generalizable conclusions about the superiority of one approach over another can be made.

Reproducibility Probability Score—incorporating measurement variability across laboratories for gene selection

To the editor:

In the September issue, a paper entitled "The MicroArray Quality Control (MAQC) project shows interplatform reproducibility of gene expression measurements" (Shi, L. *et al., Nat. Biotechnol.* **24**, 1151–1161, 2006) authored by us and others highlighted the need for a statistical metric to account for interlaboratory measurement variability in the selection of differentially expressed genes from microarray data. Here, we describe a novel metric (**Supplementary Methods 1** online) called the Reproducibility Probability Score (RPS), which is computed from gene expression data from a single laboratory. A gene with a higher RPS is evidence for differential expression that is more reproducible by other laboratories. We also provide a free, open source program (http://biocomp.bioen.uiuc.edu/rps) for computing the RPS to identify differentially expressed genes. Currently, the RPS program is capable of analyzing data from five commonly used microarray platforms—the Human Genome Survey Microarray v2.0 (Applied Biosystems, Foster City, CA, USA), the HG-U133 Plus 2.0 GeneChip (Affymetrix, Santa Clara, CA, USA), the Whole Human Genome Oligo Microarray G4112A (Agilent, Palo Alto, CA, USA), the CodeLink Human Whole Genome (GE Healthcare, Chalfont St. Giles, UK) and the Human-6 BeadChip 48K v1.0 (Illumina, San Diego)—and it can be extended to analyze other microarray platforms.

The RPS for a gene is defined as the probability that this gene is selected as being differentially expressed from the data generated by a typical laboratory. A typical laboratory is either the original laboratory that generated the microarray data or a hypothetical laboratory that prudently follows the same protocol to study the same biological materials as the original laboratory. To compute the RPS for a gene, the user needs to choose a traditional gene selection procedure, denoted by $\Theta > \theta$, where Θ is a statistic, such as the *P* value (or its inverse, if > in the gene selection procedure is interpreted literally), or a set of statistics, and θ is its corresponding threshold(s). The RPS for a gene is:

$$\text{RPS} = \textit{Prob}\ (\text{this gene is selected by a typical laboratory}) = E_k\{I[\Theta_k > \theta]\} \quad (1)$$

where k is the index of typical laboratories ($k = 0,1,2,\ldots$), with $k = 0$ denoting the original laboratory and $k > 0$ denoting the hypothetical laboratories. $E_k\{\bullet\}$ is the expectation over k. $I[\bullet]$ is the 0–1 indicator function, and Θ_k is the user-chosen metric computed from the data from the k^{th} laboratory. If there were a perfect correlation in the interlaboratory measurements, equation (1) would reduce into:

$$\text{RPS} = I[\Theta_0 > \theta] \quad (2)$$

which is identical to gene selection based on the user-specified procedure on the data from the original laboratory.

The RPS program uses simulation to generate data for the hypothetical laboratories in the computation of the RPS. Two sets of data are used in the simulation. The first set is the data generated from the original laboratory, also referred to as the new data. The second set is a reference data set. Currently, the RPS program uses the data from the Shi *et al.* as the reference data set, and it provides an option for the use of other reference data. The reference data and the new data have to be generated on the same microarray platform, and the reference data have to be generated by more than one laboratory. However, the reference data do not have to be generated by the same laboratory or originate from the same biological samples as the new data.

The workflow of the RPS program is as follows (**Fig. 1** and **Supplementary Methods 2** online). First, the RPS program applies a mixed-effects model to the reference data to estimate interlaboratory correlation for each probe (set). By default the MAQC data are preloaded as the reference data, but if the new data to be analyzed come from a microarray platform with no preloaded data, the user should provide appropriate reference data. Second, the RPS program reads in the new data to be analyzed. It uses the new data together with the estimated interlaboratory correlations to simulate the data for the hypothetical laboratories. Finally, the program uses the new data and the simulated data to compute an RPS for each gene, and ranks the genes by their RPS values.

To demonstrate the merit of the RPS algorithm, we have generated microarray data from five colorectal adenocarcinomas and matched normal colonic tissues. The RNA was first hybridized onto Affymetrix HG-U133-Plus-2.0 arrays in the Stanford Genome Technology Center (SGTC; Palo Alto, CA, USA). The RPS program and some other commonly used programs or procedures were used to identify differentially expressed genes. The other programs and procedures included MAANOVA[1], BayesAnova[2], FDR (Benjamini & Hocheberg procedure[3] on the two sample *t*-test), *P* value from the two sample *t*-test (computed by dChip[4]) and fold-change.

The same biological materials were then processed for hybridization in a different laboratory (the PAN facility on the Stanford