Published in final edited form as:

Ann N Y Acad Sci. 2012 July ; 1260: 95–100. doi:10.1111/j.1749-6632.2011.06383.x.

Accurate evaluation and analysis of functional genomics data and methods

Casey S. Greene¹ and Olga G. Troyanskaya^{1,2}

¹Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey

²Department of Computer Science, Princeton University, Princeton, New Jersey

Abstract

The development of technology capable of inexpensively performing large-scale measurements of biological systems has generated a wealth of data. Integrative analysis of these data holds the promise of uncovering gene function, regulation, and, in the longer run, understanding complex disease. However, their analysis has proved very challenging, as it is difficult to quickly and effectively assess the relevance and accuracy of these data for individual biological questions. Here, we identify biases that present challenges for the assessment of functional genomics data and methods. We then discuss evaluation methods that, taken together, begin to address these issues. We also argue that the funding of systematic data-driven experiments and of high-quality curation efforts will further improve evaluation metrics so that they more-accurately assess functional genomics data and methods. Such metrics will allow researchers in the field of functional genomics to continue to answer important biological questions in a data-driven manner.

Keywords

bioinformatics; functional genomics; evaluation

Introduction

With the past decade's explosion of genome-scale biomedical data, innovative computing approaches in genomics can revolutionize our understanding of biology and medicine and provide unprecedented levels of public access to biomedical information.¹ Genomics experiments now not only provide a "parts list" of the organism in the form of its genomic sequence, but also can assess what these genes do, how they are controlled in cellular pathways, and what malfunctions in these cellular systems lead to disease.²

However, despite this explosion of high-throughput data in molecular biology, our functional understanding of cellular processes remains incomplete. This gap between data generation and the discovery of reliable functional information is largely due to the lack of

Address for correspondence: Olga G. Troyanskaya, Lewis-Sigler Institute for Integrative Genomics, 242 Carl Icahn Laboratory, Princeton University, Princeton, NJ 08544. ogt@genomics.princeton.edu.

Conflicts of interest

The authors declare no conflicts of interest.

specificity and resolution in high-throughput data. In other words, the challenge lies in identifying the true biological signal and separating it from both technical and experimental noise. To alleviate this lack of specificity and extract accurate functional information from these data, integrated analysis of heterogeneous data sources with robust computational methods is necessary.³ Ideally, such analyses can leverage multiple, often complementary, sources of information in diverse biological datasets to generate precise biological hypotheses about what proteins do (gene function),^{4–6} how they do it (interactions and regulation),^{7,8} and how various perturbations (i.e., disease or drugs) affect biological pathways.^{9,10}

The noisy and heterogeneous nature of functional genomic data and the lack of accurate gold standards, or biological “truths” that can be used to assess accuracy and coverage of each dataset, make it challenging to analyze these data and generate specific, experimentally testable hypotheses. Even the best algorithms can be stymied by significant evaluation biases that can lead to trivial or incorrect predictions with apparently higher accuracy. These biases are often highly technical and data dependent, yet their understanding is critical to any analysis of these diverse functional genomics data, be it for prediction of protein function and interactions, or for more complex modeling tasks, such as building biological pathways. Below, we discuss the major sources of these biases, and propose ways to detect and avoid them when dealing with functional genomics datasets. When these biases are properly addressed, functional genomics approaches have great potential to make high-quality predictions of gene function, disease involvement, or tissue specificity. These predictions can be used to efficiently and effectively target genes for further experimental study.

Biases in analysis and evaluation of functional genomics data

The biases that affect the analysis and evaluation of functional genomics data can arise from biological factors and from the actual process of measuring biological systems. Here, we discuss four such biases, which we call process bias, term bias, standard bias, and annotation distribution bias.

Process bias can arise when distinct biological groups of genes, processes, or functions, such as the cell cycle DNA damage repair, are grouped for evaluation. A classic example of this comes from the ribosome pathway in yeast. Ribosomes are the organelles in the cell responsible for translating mRNA to protein, and their components are very easily measured by a specific experimental platform called gene expression microarrays, which comprise the majority of existing functional genomics data. Myers *et al.*¹¹ showed that, when evaluation is performed on the grouped set of processes, a single easy-to-predict process, such as the ribosome, can dramatically alter the evaluation results (Fig. 1). In the case of the ribosome, this difference is so pronounced that the gene coexpression data, which is clearly the best overall predictor when the ribosome is grouped with other processes, is only of average utility once the ribosome pathway is removed from a set of 99 pathways used for evaluation.

Term bias arises when the gold standard is correlated with other factors. At its most extreme, this can occur because of the direct inclusion of evaluation standards in data used for

training, such as inclusion of one curated database (for example, Gene Ontology)^{12,13} as input data when another (for example, KEGG)¹⁴ is used for evaluation. In functional genomics data, these term biases can also occur through subtle contamination. One example of this hard-to-detect yet significant bias comes from the presence or absence of genes in large-scale genomic datasets. Because genes in most large-scale datasets are not randomly selected (for example, genes measured by microarray platforms), genes with more associated terms are more likely to be well represented. To a researcher who assumes that the presence or absence of genes in the data is random, using the sum of a score for a gene would not alter performance from using the mean score. In practice, however, there is a term bias related to which genes are well represented. Therefore, the sum appears dramatically better than the mean, but this is only because the sum acts as a proxy for how well studied the gene is. In this way, small algorithmic differences can lead to large performance changes when they exploit subtle, or even hidden, contamination between the training and evaluation sets.

Standard bias can impact evaluation methods because, in the traditional biological literature, biologists do not randomly select genes for the study. Researchers studying a specific phenotype will select genes that they expect to be involved in the process for assay. This selection step is highly non-random, and can create an unexpected and difficult to assess bias. Huttenhower *et al.*¹⁵ found that in the case of mitochondrial biogenesis, while methods' ability to predict gene-phenotype relationships assessed through cross-validation differed, the actual ability of the methods to predict novel relationships (verified through experimental study) was very similar (Fig. 2). Furthermore, Hibbs *et al.*¹⁶ observed that existing annotations of genes to phenotypes are often biased toward very severe gene-phenotype relationships and that subtle phenotypes are underrepresented. These two results, taken together, show how the process by which experiments are performed in molecular biology and genetics can introduce subtle biases into standards that mislead even very careful evaluation. These results also highlight the potential of functional genomics. By providing high-quality predictions from large-scale data compendia, a targeted experimental study based on carefully designed functional genomics approaches can reduce the bias in our existing knowledge.

The final challenge that we discuss is *annotation distribution bias*. This occurs because genes are not evenly annotated to functions and phenotypes. This broad size distribution creates difficulties in the assessment of functional genomics data and methods.¹⁷ The best predictions for biological follow-up are both highly specific and accurate, but the best predictions from a performance standpoint (as performance is generally measured) are very broad functions or phenotypes. This is because predictions to these broad terms are more likely to be accurate by chance alone. We have set up a web server at <http://dfp.princeton.edu> that illustrates this effect. The web server predicts future annotations to the Gene Ontology (GO),^{12,13} a common source of functional annotations, based on existing annotations as of an "evaluation cutoff" date. The predictions are made solely based on the number of existing annotations to each term, and thus use no gene-specific information (in other words, this is "data-free prediction"). The performance of this method is competitive with modern prediction methodologies.⁵ This indicates that performance on standard metrics

does not necessarily reflect the actual utility of functional genomics data or methods, necessitating careful correction for such biases.

Avoiding and addressing evaluation biases

Despite these challenges, it is possible to perform meaningful evaluation of functional genomics data and methods through careful and critical assessment. There are computational solutions that, when used with care, address these challenges and make accurate and unbiased evaluation possible. There are also ways to effectively use additional experimental data to supplement computation analyses. Finally, computationally directed, comprehensive experimental follow-up can be performed. Such systematically directed experiments are the ideal, though often too costly in time and money, solution, as they address all of the above concerns by providing direct experimental confirmation of results.

Individual computational assessments each have weaknesses that can compromise their utility, but used carefully and taken together they can provide a great deal of information about the effectiveness of functional genomics methods and data. To avoid process bias, distinct processes should be evaluated separately. If a single summary statistic is required, distinct functions should only be combined after insuring there are no outliers that will dramatically change the interpretation of results. If there are outliers, results should be provided with and without outliers, as in Myers *et al.*¹¹

Term bias can be mitigated with the addition of a temporal holdout. As long as the functional genomics data are fixed on a certain day, and phenotype or function assignments after that day are used for evaluation, this metric should avoid many of the hidden circularity issues that can affect simple random holdouts. Many functional genomics studies use either a temporal or a random holdout,^{18–23} but using both can help mitigate the common biases that we identify. It is also important to realize that not all correct predictions are uniformly useful. Many easy-to-predict terms are highly generic (for example protein phosphorylation), yet good evaluation numbers can be obtained relatively easily from such not-so-useful predictions. Instead, the analysis of functional genomics data and methods should take into account the utility of predictions, either through the annotation of expert biologists¹¹ or metrics of prediction specificity.²⁴

Standard bias is the most difficult to address. One way of avoiding standard bias is by making a final performance assessment through a blinded literature review. Because there are not sufficient professional curators to fully catalog all experimental results for gold standard datasets, often unannotated examples still remain in the literature. Data-driven methods can then discover the same phenotype, function, or tissue relationship. A careful analysis of the literature can then discover these underannotations, which can be used as a benchmark in a hand-curated or even in an automated manner. For each function of interest, genes from the function are paired with randomly selected genes. This set of genes is shuffled and provided in an unlabeled form. These genes are then evaluated for their role in the function of interest based on currently available literature. After genes are classified based by literature evidence, the frequency of literature support for predictions from the

method or data to those chosen randomly provides an assessment of the quality of the data or method.

Of course, the most definitive way to assess functional genomics data and methods is by the biological validation of predictions. This can be done as part of a predefined experimental pipeline. For example, Hess *et al.* predicted and validated more than 100 proteins that affect the biogenesis and inheritance of mitochondria in *Saccharomyces cerevisiae*.²⁵ Interestingly, the computationally directed experimental approach allowed for the discovery of more genes with subtle mitochondrial phenotypes than had previously been observed.¹⁶ Practical and cost concerns dictate that this can only occur in a small subset of studies, but when feasible, this provides the highest quality assessment. This evaluation can also be done on a large multilab scale when sufficient funding is available. The COMBEX project now provides grants to biologists that agree to test prespecified computational predictions.²⁶ To insure the utility of such work, the evaluation data from such experiments (including both confirmed and unconfirmed predictions) should be freely shared with the community to provide high-quality evaluation metrics for other studies.

Discussion

The difficulty of evaluating functional genomics data presents distinct challenges to the fields of functional genomics, but there are challenges that can be addressed through careful experimental design. The community needs to be very consistent and careful during evaluation to insure that neither those developing nor those using the algorithms or data are misled by incomplete evaluations. Publicly available gold standard sets derived from systematic high-quality experiments, such as those now available in mitochondria,²⁵ are very helpful for evaluation, but more are required. The COMBEX strategy of prioritizing experiments that evaluate functional genomics predictions may also help in this regard.²⁶

Current standard computational evaluation methods often rely upon curated GO¹³ or KEGG¹⁴ annotations. These functional annotations are extremely useful in this regard, but are incomplete due to the breadth of modern biomedical and biological literature. More funding allocated for such expert curation is a cost-effective way to convert information from biomedical literature into well-structured knowledge. Expanding funding for human curators will expand the coverage of GO and other such resources, thereby broadening the applicability of evaluation metrics that use these sources. A complementary strategy to increase the proportion of structured knowledge uses structured publication^{27,28} and next-generation bioinformatics tools like PILGRM²⁹ that allow biologists to direct the analysis of large data compendia. Because biologists use their own literature review to direct the analysis, they effectively curate a source of structured knowledge suitable for evaluation. In combination, these approaches will support more accurate and useful predictions, thus enabling faster biological discovery by providing highly specific and unbiased biological hypotheses for experimental testing.

References

1. Hieter P, Boguski M. Functional genomics: it's all how you read it. *Science*. 1997; 278:601–602. [PubMed: 9381168]

2. Vaske CJ, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*. 2010; 26:i237–i245. [PubMed: 20529912]
3. Louie B, Mork P, Martin-Sanchez F, et al. Data integration and genomic medicine. *J Biomed Inform*. 2007; 40:5–16. [PubMed: 16574494]
4. Fraser AG, Marcotte EM. A probabilistic view of gene function. *Nat Genet*. 2004; 36:559–564. [PubMed: 15167932]
5. Pena-Castillo L, et al. A critical assessment of *Mus musculus* gene function prediction using integrated genomic evidence. *Genome Biol*. 2008; 9(Suppl 1):S2. [PubMed: 18613946]
6. Mostafavi S, Ray D, Warde-Farley D, et al. Gene-MANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol*. 2008; 9(Suppl 1):S4. [PubMed: 18613948]
7. Hart GT, Lee I, Marcotte ER. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics*. 2007; 8:236. [PubMed: 17605818]
8. Warde-Farley D, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res*. 2010; 38:W214–W220. [PubMed: 20576703]
9. Setlur SR, et al. Integrative microarray analysis of pathways dysregulated in metastatic prostate cancer. *Cancer Res*. 2007; 67:10296–10303. [PubMed: 17974971]
10. Lee I, et al. A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans*. *Nat Genet*. 2008; 40:181–188. [PubMed: 18223650]
11. Myers CL, Barrett DR, Hibbs MA, et al. Finding function: evaluation methods for functional genomic data. *BMC Genomics*. 2006; 7:187. [PubMed: 16869964]
12. Ashburner M, et al. Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat Genet*. 2000; 25:25–29. [PubMed: 10802651]
13. Reference Genome Group of the Gene Ontology Consortium. The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. *PLoS Comput Biol*. 2009; 5:e1000431. [PubMed: 19578431]
14. Kanehisa M, et al. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 1999; 27:29–34. [PubMed: 9847135]
15. Huttenhower C, et al. The impact of incomplete knowledge on evaluation: an experimental benchmark for protein function prediction. *Bioinformatics*. 2009; 25:2404–2410. [PubMed: 19561015]
16. Hibbs MA, et al. Directing experimental biology: a case study in mitochondrial biogenesis. *PLoS Comput Biol*. 2009; 5:e1000322. [PubMed: 19300515]
17. Gillis J, Pavlidis P. The impact of multifunctional genes on “guilt by association” analysis. *PLoS One*. 2011; 6:e17258. [PubMed: 21364756]
18. King RD, Karwath A, Clare A, et al. Accurate prediction of protein functional class from sequence in the *Mycobacterium tuberculosis* and *Escherichia coli* genomes using data mining. *Yeast*. 2000; 17:283–293. [PubMed: 11119305]
19. Kasif S, et al. Whole-genome annotation by using evidence integration in functional-linkage networks. *P Natl Acad Sci USA*. 2004; 101:2888–2893.
20. Massjouni N, Rivera CG, Murali TM. VIRGO: computational prediction of gene functions. *Nucleic Acids Res*. 2006; 34:W340–W344. [PubMed: 16845022]
21. Chen, Y.; Li, Z.; Liu, J. Learning Kernel Matrix from Gene Ontology and Annotation Data for Protein Function Prediction *Advances in Neural Networks–ISNN 2009*. Yu, W.; He, H.; Zhang, N., editors. Vol. 5553. Springer; Berlin/Heidelberg: 2009. p. 694-703.
22. Lippert C, Ghahramani Z, Borgwardt KM. Gene function prediction from synthetic lethality networks via ranking on demand. *Bioinformatics*. 2010; 26:912–918. [PubMed: 20154010]
23. Kourmpetis YA, van Dijk AD, Bink MC, et al. Bayesian Markov Random Field analysis for protein function prediction based on network data. *PLoS One*. 2010; 5:e9293. [PubMed: 20195360]

24. Lord PW, Stevens RD, Brass A, et al. Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics*. 2003; 19:1275–1283. [PubMed: 12835272]
25. Hess DC, et al. Computationally driven, quantitative experiments discover genes required for mitochondrial biogenesis. *PLoS Genet*. 2009; 5:e1000407. [PubMed: 19300474]
26. Kasif S, et al. COMBREX: a project to accelerate the functional annotation of prokaryotic genomes. *Nucleic Acids Res*. 2011; 39:D11–D14. [PubMed: 21097892]
27. Gerstein M, Seringhaus M, Fields S. Structured digital abstract makes text mining easy. *Nature*. 2007; 447:142. [PubMed: 17495904]
28. Ripple AM, Mork JG, Knecht LS, et al. A retrospective cohort study of structured abstracts in MEDLINE, 1992–2006. *J Med Libr Assoc*. 2011; 99:160–163. [PubMed: 21464855]
29. Greene CS, Troyanskaya OG. PILGRM: an interactive data-driven discovery platform for expert biologists. *Nucleic Acids Res*. 2011; 39:W368–W374. [PubMed: 21653547]

Figure 1.

The performance of microarray coexpression data (the blue line) appears to be by far the best predictor (quality is indicated by precision at the specified number of true positive pairs) when the ribosome is included among the evaluation pathways (triangle symbols). When the ribosome is removed from consideration (circle symbols), coexpression is not substantially better than the other types of data.

Figure 2.

The expected AUPRC from cross-validation (diagonally hashed) and actual performance during experimental validation (solid) of various functional genomics strategies for gene-function prediction. The fold change between the expected and actual performance of the integrative method (bioPIXIE) is dramatically different than that for the coexpression methods MEFIT and SPELL. The expected performance can be influenced by subtle biases, so careful analysis is required to insure that results will generalize to actual performance.