Open Access

Introduction

A race through the maze of genomic evidence

Timothy R Hughes* and Frederick P Roth†

Addresses: *Donnelly Centre for Cellular and Biomolecular Research and Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario M5S3E1, Canada. †Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School and Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA.

Correspondence: Timothy R Hughes. Email: t.hughes@utoronto.ca. Frederick P Roth. Email: fritz_roth@hms.harvard.edu

Published: 27 June 2008

Genome ***Biology*** 2008, **9:**S1

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2008/9/S1/S1

One of the most surprising aspects of the completed human and mouse genome sequences [1-3] has been the relatively small number of protein-coding genes. The current estimate of <24,000 protein-coding genes in human and mouse is only four times that of budding yeast [4]. A complete encyclopedia of biochemical, cellular, and physiological gene functions is now an immediate rather than a long-term goal.

The unifying theme for papers in this supplement to *Genome Biology* is the automated inference of molecular function of gene products, and of their membership within cellular components and biological processes. In each study, thousands of variables describing genes and gene-gene relationships have been integrated using machine learning methods to infer Gene Ontology (GO) terms for essentially all genes in the studied genome.

Systematic biochemical and genetic experimentation in simpler model organisms has contributed to the rapid increase in the proportion of characterized genes. For example, about 80% of yeast genes have some annotated function [5]. Beyond simpler model systems such as yeast, however, cost and time requirements may preclude many systematic analyses, for example, resource-intensive phenotype assays in adult animals. Fortunately, efforts in simpler model organisms have illustrated that gene functions can be inferred on the basis of data types that are easier to collect systematically. Protein sequence features, expression patterns, and protein-protein interactions, for example, can provide powerful clues to function. This raises the prospect of directing resource-intensive experimentation toward the genes most likely to yield positive results. In yeast, this concept is well established, and the tradeoffs between performance measures, the efficacy of combinations of different data types in making different types of predictions, and the applicability of diverse inference algorithms are topics of active research.

Large scale experimentation in mammals is coming of age. A wide variety of mRNA expression analysis experiments are available in public data repositories, for example, the Gene Expression Omnibus [6]. The majority of human genes have at least one moveable open reading frame clone [7-9], enabling expression studies *in vitro* and in model systems. 'Knockdown' reagents targeting most mouse and human genes are now available [10], facilitating analysis of biochemical and cellular gene functions. Efforts are underway to create a mutant allele of each mouse gene [11], which will enable analysis of physiological and developmental roles.

The first paper in this supplemental issue [12] describes the 'MouseFunc' challenge, in which nine bioinformatics teams independently predicted mouse GO terms. Importantly, each used a common collection of training data and common benchmarks, which allowed comparison among the inference methods, data sets, and categories of gene functions. Predictions were tested using cross-validation (annotation for a subset of genes was hidden from the participants). Predictions were further tested by two forms of prospective evaluation: first, using GO annotations that had been added to the database since the inception of the study; and second, literature related to top-scoring novel predictions was investigated intensively by experienced mouse biologists.

Each of the companion papers in this issue is connected to the MouseFunc challenge, either in the nature of the algorithms used, in the datasets employed, or both. Guan and coworkers (led by Olga Troyanska) apply a support vector machine approach to predict mouse gene function [13]. They go on to apply this approach to the more tractable model eukaryote *Saccharomyces cerevisiae* and to test specific predictions experimentally. Mostafavi and colleagues (led by Quaid Morris) apply a ridge regression approach to predict mouse and yeast gene function [14]. The approach is quite fast, permitting their 'GeneMania' software to perform predictions 'on the fly' with a training set provided by the user. Kim and coworkers (led by Edward Marcotte) infer mouse gene function both directly and via a functional linkage network [15]. Functional linkage graphs contain connections between genes weighted by confidence that are functionally related [16]. Obozinski and colleagues (led by William Noble) investigate the possibility of inconsistency between predictions of different functions [17]. For example, it is possible for some approaches to assign a higher prediction score to 'DNA helicase activity' than to its logical parent term 'helicase activity'. They show that 'reconciliation' methods that enforce consistency between different GO term predictions can improve performance. Tian and coworkers [18] and Tasan and colleagues [19], teams both led by one of us (FPR), each combine guilt-by-profiling and guilt-by-association approaches to make predictions. Tian and coworkers describe the methodology and apply it to predict *S. cerevisiae* gene functions, while Tasan and colleagues apply the methodology to predict both functions and phenotypes for mouse genes.

Many other quantitative fields have benefitted by standardization of training and test sets. For example, the Critical Assessment of Techniques for Protein Structure Prediction (CASP) challenge [20] has made rigorous comparisons among protein structure predictions. This special issue suggests the value of similar standardization in the arena of function prediction.

Importantly, inferences about function and phenotype made in this issue are not black or white, but rather are expressed in shades of gray. Biology will long remain in the 'working model' phase, in which each statement about a gene's role must be accompanied by some uncertainty. An honest assessment of our uncertainties could allow us to direct resources efficiently to those experiments most likely to resolve these uncertainties. Quantitative predictions allow individual users requiring highly stringent predictions to impose a high prediction score threshold, while users may lower their threshold and include additional false positives if they wish to cast a wide net and catch more true positives.

The approaches taken in this issue have common limitations. To reduce the scope of the computational problem and eliminate the potential for inflated performance estimates due to circular reasoning, participants did not have access to GO annotations from other species. Although the training data did incorporate many previous transfers of annotation from other species by orthology, these methods could also benefit from a similar standardization and benchmarking strategy.

We also note that identifying the best strategies does not always help us to understand why the best strategies worked well. Because of the computationally intensive nature of function prediction, only a limited number of variant approaches were evaluated. A full factorial analysis of variations on the most successful strategies will help provide this understanding and allow future optimization.

The high precision of top predictions for many GO terms illustrates the richness and value of data sources that have become available for mammals over recent years. However, one lesson learned is that it is difficult to achieve both high precision and high recall. Currently, no algorithms achieve both for most functional categories. Improvements in either the inference methods, the problem setup, or in the information content of the data sets themselves will be needed in order to make a major dent in the more than 10,000 currently uncharacterized mouse and human genes.

Abbreviations

GO, Gene Ontology.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This article has been published as part of *Genome Biology* Volume 9 Supplement 1, 2008: Quantitative inference of gene function from diverse large-scale datasets. The full contents of the supplement are available online at http://genomebiology.com/supplements/9/S1

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, *et al.*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, *et al.*: **The sequence of the human genome.** *Science* 2001, **291:**1304-1351.
3. Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, *et al.*: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420:**520-562.
4. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H,

Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG: **Life with 6000 genes.** *Science* 1996, **274:**563-547.
5. Pena-Castillo L, Hughes TR: **Why are there still over 1000 uncharacterized yeast genes?** *Genetics* 2007, **176:**7-14.
6. Barrett T, Edgar R: **Gene Expression Omnibus: microarray data storage, submission, retrieval, and analysis.** *Methods Enzymol* 2006, **411:**352-369.
7. Lamesch P, Li N, Milstein S, Fan C, Hao T, Szabo G, Hu Z, Venkatesan K, Bethel G, Martin P, Rogers J, Lawlor S, McLaren S, Dricot A, Borick H, Cusick ME, Vandenhaute J, Dunham I, Hill DE, Vidal M: **hORFeome v3.1: a resource of human open reading frames representing over 10,000 human genes.** *Genomics* 2007, **89:**307-315.
8. Nagase T, Yamakawa H, Tadokoro S, Nakajima D, Inoue S, Yamaguchi K, Itokawa Y, Kikuno RF, Koga H, Ohara O: **Exploration of human ORFeome: high-throughput preparation of ORF clones and efficient characterization of their protein products.** *DNA Res* 2008 in press.
9. Bechtel S, Rosenfelder H, Duda A, Schmidt CP, Ernst U, Wellenreuther R, Mehrle A, Schuster C, Bahr A, Blöcker H, Heubner D, Hoerlein A, Michel G, Wedler H, Köhrer K, Ottenwälder B, Poustka A, Wiemann S, Schupp I: **The full-ORF clone resource of the German cDNA Consortium.** *BMC Genomics* 2007, **8:**399.
10. Silva JM, Li MZ, Chang K, Ge W, Golding MC, Rickles RJ, Siolas D, Hu G, Paddison PJ, Schlabach MR, Sheth N, Bradshaw J, Burchard J, Kulkarni A, Cavet G, Sachidanandam R, McCombie WR, Cleary MA, Elledge SJ, Hannon GJ: **Second-generation shRNA libraries covering the mouse and human genomes.** *Nat Genet* 2005, **37:**1281-1288.
11. Collins FS, Rossant J, Wurst W: **A mouse for all reasons.** *Cell* 2007, **128:**9-13.
12. Peña-Castillo L, Tasan M, Myers CL, Lee H, Joshi T, Zhang C, Guan Y, Leone M, Pagnani A, Kim WK, Krumpelman C, Tian W, Obozinski G, Qi Y, Mostafavi S, Lin GN, Berriz GF, Gibbons FD, Lanckriet G, Qiu J, Grant C, Barutcuoglu Z, Hill DP, Warde-Farley D, Grouios C, Ray D, Blake JA, Deng M, Jordan MI, Noble WS, *et al.*: **A critical assessment of *Mus musculus* gene function prediction using integrated genomic evidence.** *Genome Biol* 2008, **9(suppl 1):**S2.
13. Guan Y, Myers CL, Hess DC, Barutcuoglu Z, Caudy AA, Troyanskaya OG: **Predicting gene function in a hierarchical context with an ensemble of classifiers.** *Genome Biol* 2008, **9(suppl 1):**S3.
14. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q: **GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function.** *Genome Biol* 2008, **9(suppl 1):**S4.
15. Kim WK, Krumpelman C, Marcotte EM: **Inferring mouse gene functions from genomic-scale data using a combined functional network/classification strategy.** *Genome Biol* 2008, **9(suppl 1):**S5.
16. Lee I, Date SV, Adai AT, Marcotte EM: **A probabilistic functional network of yeast genes.** *Science* 2004, **306:**1555-1558.
17. Obozinski G, Lanckriet G, Grant C, Jordan MI, Stafford Noble W: **Consistent probabilistic outputs for protein function prediction.** *Genome Biol* 2008, **9(suppl 1):**S6.
18. Tian W, Zhang LV, Taşan M, Gibbons FD, King OD, Park J, Wunderlich Z, Cherry JM, Roth FP: **Combining guilt-by-association and guilt-by-profiling to predict *Saccharomyces cerevisiae* gene function.** *Genome Biol* 2008, **9(Suppl 1):**S7.
19. Taşan M, Tian W, Hill DP, Gibbons FD, Blake JA, Roth FP: **An *en masse* phenotype and function prediction system for *Mus musculus*.** *Genome Biol* 2008, **9(suppl 1):**S8.
20. Moult J: **Rigorous performance evaluation in protein structure modelling and implications for computational biology.** *Phil Trans R Soc Lond B Biol Sci* 2006, **361:**453-458.