W340–W344 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl225

VIRGO: computational prediction of gene functions

Naveed Massjouni, Corban G. Rivera and T. M. Murali*

660 McBryde Hall, Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg VA 24061, USA

Received February 14, 2006; Revised March 25, 2006; Accepted March 27, 2006

ABSTRACT

Dramatic advances in sequencing technology and sophisticated experimental assays that interrogate the cell, combined with the public availability of the resulting data, herald the era of systems biology. However, the biological functions of more than 40% of the genes in sequenced genomes are unknown, posing a fundamental barrier to progress in systems biology. The large scale and diversity of available data requires the development of techniques that can automatically utilize these datasets to make quantified and robust predictions of gene function that can be experimentally verified. We present a service called the VIRtual Gene Ontology (VIRGO) that (i) constructs a functional linkage network (FLN) from gene expression and molecular interaction data, (ii) labels genes in the FLN with their functional annotations in the Gene Ontology and (iii) systematically propagates these labels across the FLN in order to precisely predict the functions of unlabelled genes. VIRGO assigns confidence estimates to predicted functions so that a biologist can prioritize predictions for further experimental study. For each prediction, VIRGO also provides an informative 'propagation diagram' that traces the flow of information in the FLN that led to the prediction. VIRGO is available at http://whipple.cs.vt.edu:8080/virgo.

MOTIVATION

More than 250 complete genome sequences are now available, including those of 35 eukaryotes (1). Increasingly sophisticated high-throughput biological experiments provide a wide range of functional genomic information about cell state. These advances, combined with the public availability of these datasets, herald the era of systems biology (2,3). However, a fundamental roadblock to progress in systems biology is the poor state of knowledge about the biological functions of the genes in sequenced genomes (4,5). Using sequence similarity to predict gene function provides annotations only for about 40% of eukaryotic genes (6). Some of these annotations may also be incorrect, as they are transmitted from one genome to another via weak chains of inference (7). Genes of unknown function might support important cellular functions. Discovering the functions of these genes will provide critical insights into the biology of many organisms. In addition, discovering these functions will improve our ability to annotate genomes sequenced in the future. The large scale and diversity of available functional genomic data requires the development of novel computational tools that can automatically integrate these data in order to compute quantified and testable predictions of the functions of poorly understood genes.

In this paper, we provide a powerful interface called 'the VIRtual Gene Ontology' (VIRGO) that enables a biologist to

(i) integrate gene expression data collected in the laboratory with molecular interaction networks,
(ii) construct a functional linkage network (FLN) from these datasets,
(iii) label the genes in the FLN with functional annotations from the Gene Ontology (GO) (8) and
(iv) systematically propagate these labels across the FLN in order to predict the functions of unlabelled genes.

The biologist can query VIRGO for predictions of interest and prioritize them using confidence values assigned by VIRGO. VIRGO also provides informative 'propagation diagrams' that trace the flow of information in the FLN. These diagrams may assist the biologist in ascertaining the rationale behind a prediction. A number of powerful methods have been published for predicting gene function by integrating different types of functional genomic data (9–16). As far as we are aware, VIRGO is the first web server that makes such a prediction engine widely available.

THE VIRGO SYSTEM

Figure 1 displays the VIRGO system. We describe its main components below.

*To whom correspondence should be addressed. Tel: +1 540 231 8534; Fax: +1 540 231 6075; Email: murali@cs.vt.edu

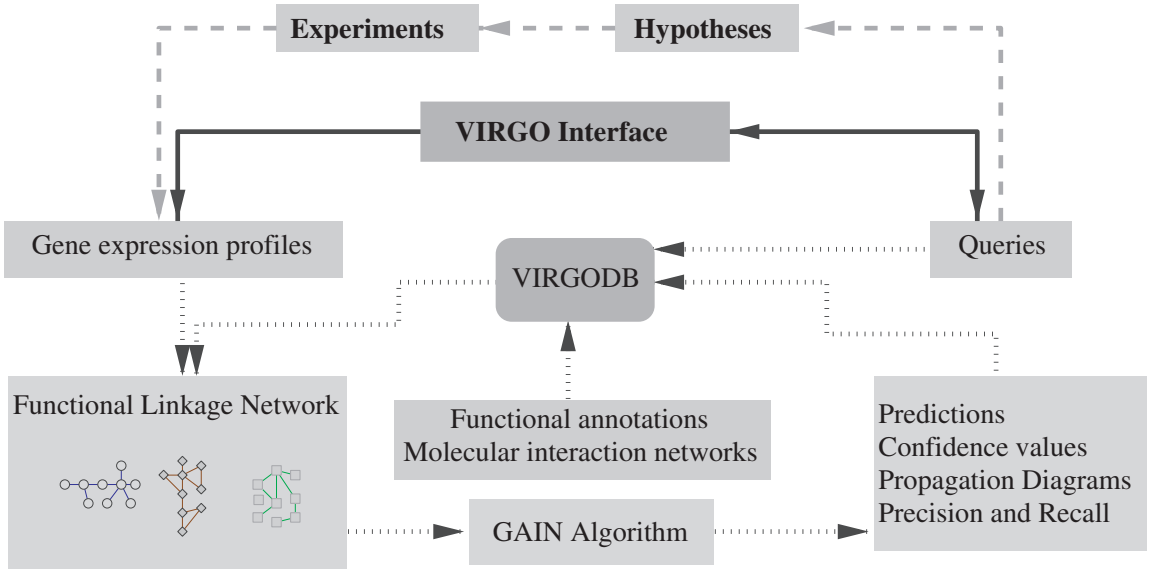

Figure 1. The VIRGO system. Solid arrows indicate a biologist's interaction with VIRGO. Dotted arrows indicate flow of information and computation within VIRGO. Dashed lines indicate generation of biological hypotheses and experimental data that we hope VIRGO will inspire.

Functional Linkage Networks

A promising basis for predicting gene function identifies functional associations of genes of unknown function with genes of known function. Diverse sources of biological data contain evidence for such associations. For instance, two genes may have the same function if their protein products interact (17,18) or if they have very similar patterns of gene expression (19,20). An FLN (21–24) is a powerful framework for representing and analysing such relationships. An FLN is a graph in which each node corresponds to a gene; the node is labelled by the set of functions that annotate the gene. An edge in an FLN connects two genes if some experimental or computational procedure suggests that these genes might share the same function. Each edge in the FLN has a real-valued weight; the sign of the weight indicates whether the connected genes share or do not share the function, while the magnitude of the weight reflects our confidence in the edge.

A number of on-line databases (24–34) have assembled large collections of functional links between genes by curating the literature or by combining multiple experimental and computational procedures. Other authors have proposed techniques for constructing FLNs that integrate multiple sources of data (22,35) or FLNs that are based on gene expression data analysed across multiple species (19,36). Although these databases and algorithms are highly valuable sources of functional associations, many of them focus on constructing FLNs and do not address the question of using FLNs for automatically predicting gene functions.

The GAIN algorithm

VIRGO uses the 'Gene Annotation using Integrated Networks' (GAIN) (21) algorithm as its function prediction engine. GAIN automatically and robustly suggests putative functions by systematically propagating functional annotations across the FLN while exploiting the constraints imposed by the topology of the FLN. In earlier work (21), we evaluated GAIN by integrating a protein–protein interaction network for *S. cerevisiae* based on the GRID database (30) and a gene expression dataset with 300 conditions (gene knockouts and chemical treatments) (37). The protein–protein interaction network provided the edges of the FLN. We assigned each edge in the FLN a weight equal to the absolute value of the Pearson's correlation coefficient of the expression profiles of the genes incident on the edge. We used the GO functional annotations for *S. cerevisiae* as of December 1, 2002. We considered those GO functions that annotated at most 10% of the genes in the FLN and for which GAIN achieved at least 75% precision and recall on average on leave-one-out cross validation. We restricted our attention to those predicted gene-function pairs where the function belonged to this set of 485 functions. Since GAIN may predict multiple functions for a gene, one predicted function may be an ancestor of another predicted function in the GO directed acyclic graph (DAG). Therefore, if GAIN predicted a gene as having two functions where one function is an ancestor of the other, we discarded the ancestor as a prediction. These steps yielded 207 predicted gene-function pairs spanning 130 distinct genes, 98 distinct functions and all three GO categories. We compared these 207 predictions to the GO annotations for *S. cerevisiae* as of March 24, 2006. We computed the distance in the GO DAG between each function predicted by GAIN for a gene (based on the 2002 dataset) and the correct annotation in the same GO category as the predicted function (if one existed in the 2006 dataset). For each gene, we selected the predicted function (in each GO category) that achieved the smallest distance to a true annotation for that gene. We calculated that 11 predictions are correct, 12 predicted functions are either parents or children of the true function in the GO DAG, 36 predicted functions are at a distance two in the GO DAG from the true function, and 3 predicted functions are at a distance three from the true function. These 62 predictions span 52 genes (GAIN predicted functions in multiple GO categories for some genes). The 78 genes involved in the remaining predictions continue to

have no biologically validated functions in the same GO category as the predicted function. A table listing all the comparisons we performed is available in the Supplementary Data. The validated predictions include nucleolus, chromatin remodeling complex, snoRNA binding RNA binding and vesicle-mediated transport. These results demonstrates GAIN's ability to make accurate predictions of gene function.

The VIRGO pipeline

VIRGO is implemented in Java 1.4 and uses the Apache Jakarta Tomcat web server. The VIRGO database uses a PostgreSQL backend. GAIN is implemented in C++. A typical session for a biologist with VIRGO involves the following steps:

(i) The biologist collects a gene expression data set in the laboratory and uploads the data to VIRGO. At this stage, the biologist has option of telling VIRGO to make the resulting predictions public, i.e. available to all users of VIRGO. VIRGO's default policy is to keep the predictions private.
(ii) VIRGO invokes GAIN to integrate the gene expression dataset with molecular interactions to construct an FLN.
(iii) GAIN processes the FLN in two separate steps. The first step uses the FLN and existing annotations in GO to compute new predictions of gene function. In the optional second step, the biologist can measure GAIN's performance using leave-one-out cross validation.
(iv) At the end of each step, VIRGO parses GAIN's output files, stores the results in the VIRGO's database and informs the biologist by email that the step has completed.
(v) The biologist queries VIRGO to find high-quality predictions using propagation diagrams, confidence estimates, and other statistics as aids. Figure 2 displays a typical propagation diagram.

In the long run, we hope that a biologist will be able to use VIRGO to develop new hypotheses and perform new experiments which will yield further datasets for analysis by VIRGO.

Supported organisms and datasets

Currently, VIRGO supports analysis for *S. cerevisiae* and *H. sapiens*. We chose these two organisms since they have large and diverse collection of protein–protein interaction datasets and gene expression measurements. We periodically download these interactions from the respective websites and functional annotations from the GO website. We use the GRID dataset (30) for *S. cerevisiae*. For *H. sapiens*, we obtained 31610 interactions between 7393 human proteins from the

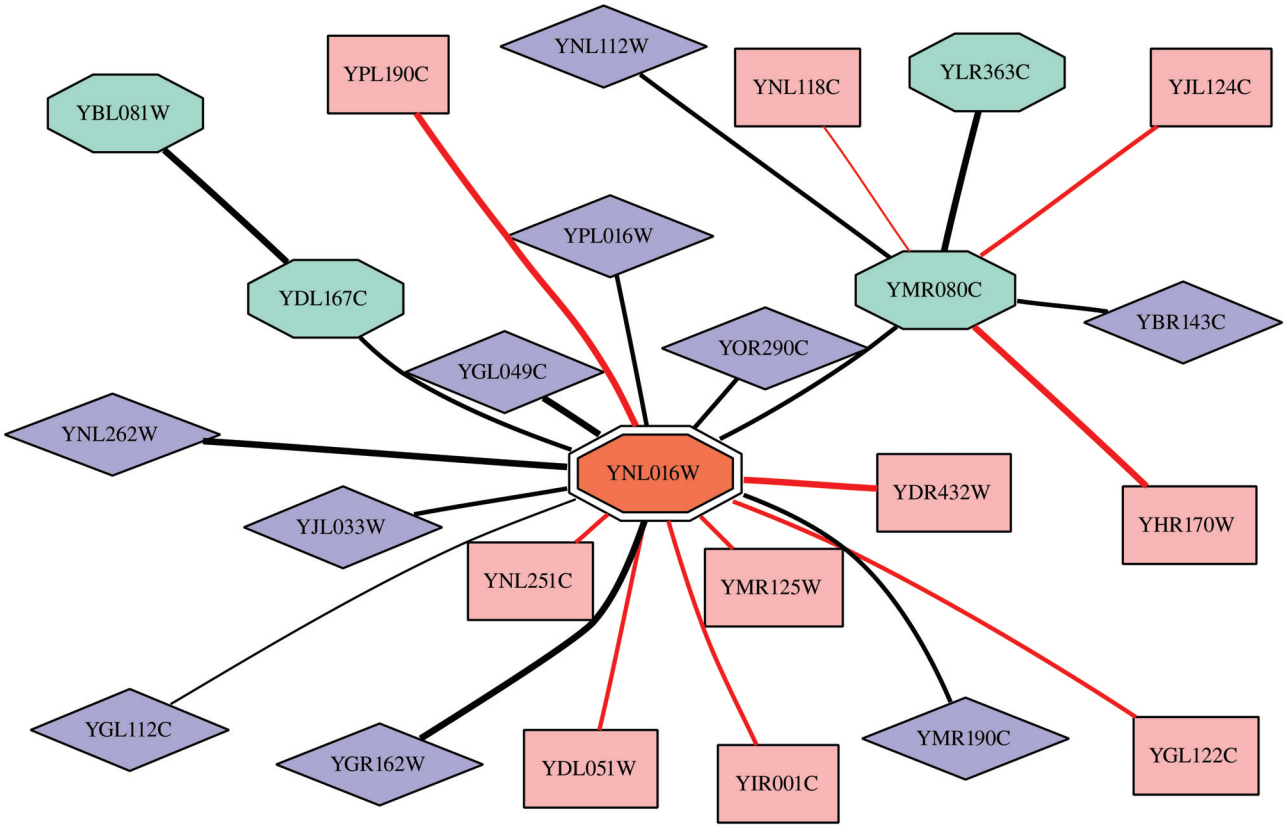

Figure 2. This propagation diagram supports the prediction that gene YNL016W (PUB1) is annotated with the biological process 'RNA binding' (GO:0000023). Red rectangles denote genes annotated with this function. Blue diamonds represent genes annotated with a different function. Octagons represent genes that either have no known function or are annotated with a function that is an ancestor of 'RNA binding.' Of these, the red octagon is the gene of interest. Other blue octagons represent genes that are also predicted to have this function. Red edges are incident on annotated nodes and help to visualize the flow of information in this network. The propagation diagram generated by VIRGO also displays edge weights, which we do not show in this picture.

IDSERVE database (29). We also included 3270 human interactions derived using large scale yeast two-hybrid experiments from Stelzl *et al*. (38), and 6726 human PPIs from Rual *et al*. (39). Overall, this human PIN contains 6274 proteins and 34087 interactions and represents interactions from a diverse variety of sources.

CONCLUSIONS AND FUTURE WORK

We have developed VIRGO, a web server for automated prediction of gene functions. A biologist can use VIRGO to obtain predictions for a system of interest by analysing relevant gene expression data integrated with molecular interactions. VIRGO provides useful auxiliary information to the biologist to assess the quality of the predictions and to prioritize them for further analysis. It is easy to extend VIRGO to other organisms for which gene expression data and functional annotations exist. We anticipate adding support for *D. melanogaster*, *C. elegans*, and *P. falciparum* and in the near future. VIRGO will also support functional predictions in organisms for which there are no publicly-available datasets of molecular interactions.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

Simon Kasif suggested the development of VIRGO. The authors thank Miguel Colon-Velez, Konstantinos Krampis, Harsha Rajasimha, Pallavi Sharma, Lachelle Waller and Andrew Warren for testing VIRGO and offering valuable feedback and suggestions. The authors thank the reviewers for numerous suggestions that improved VIRGO's functionality. Funding to pay the Open Access publication charges for this article was provided by an ASPIRES grant from the Virginia Polytechnic Institute and State University.

Conflict of interest statement. None declared.

REFERENCES

1. Bernal,A., Ear,U. and Kyrpides,N. (2001) Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res*., **29**, 126–127.
2. Ideker,T. and Lauffenburger,D. (2003) Building with a scaffold: emerging strategies for high-to low-level cellular modeling. *Trend Biotechnol*., **21**, 255–262.
3. Kitano,H. (2002) Systems biology: a brief overview. *Science*, **295**, 1662–1664.
4. Karp,P.D. (2004) Call for an enzyme genomics initiative. *Genome Biol*., **5**, 401.
5. Roberts,R.J. (2004) Identifying protein function–a call for community action. *PLoS. Biol*., **2**, E42.
6. Enright,A.J., Kunin,V. and Ouzounis,C.A. (2003) Protein families and TRIBES in genome sequence space. *Nucleic. Acids. Res*., **31**, 4632–4638.
7. Boguski,M.S. (1999) Biosequence exegesis. *Science*, **286**, 453–455.
8. Ashburner,M., Ball,C.A., Blake,J.D., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature. Genet*., **25**, 25–29.
9. Deng,M., Tu,Z., Sun,F. and Chen,T. (2004) Mapping Gene Ontology to proteins based on protein–protein interaction data. *Bioinformatics*, **20**, 895–902.
10. Lanckriet,G.R., Deng,M., Cristianini,N., Jordan,M.I. and Noble,W.S. (2004) Kernel-based data fusion and its application to protein function prediction in yeast. *Pac. Symp. Biocomput*., 300–311.
11. Mateos,A., Dopazo,J., Jansen,R., Tu,Y., Gerstein,M. and Stolovitzky,G. (2002) Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons. *Genome Res*., **12**, 1703–1715.
12. Nabieva,E., Jim,K., Agarwal,A., Chazelle,B. and Singh,M. (2005) Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, **21**, i302–i310.
13. Troyanskaya,O.G., Dolinski,K., Owen,A.B., Altman,R.B. and Botstein,D. (2003) A Bayesian frame-work forcombining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl Acad. Sci. USA*, **100**, 8348–8353.
14. Vazquez,A., Flammini,A., Maritan,A. and Vespignani,A. (2003) Global protein function prediction from protein–protein interaction networks. *Nat. Biotechnol*., **21**, 697–700.
15. Wong,S.L., Zhang,L.V. and Roth,F.P. (2005) Discovering functional relationships: biochemistry versus genetics. *Trends Genet*., **21**, 424–427.
16. Zhou,X., Kao,M.C. and Wong,W.H. (2002) Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl. Acad. Sci. USA*, **99**, 12783–12788.
17. Fields,S. and Song,O. (1989) A novel genetic system to detect protein–protein interactions. **340**, 245–246.
18. Walhout,A.J. and Vidal,M. (2001) Protein interaction maps for model organisms. *Nat. Rev. Mol. Cell. Biol*., **2**, 55–62.
19. Stuart,J.M., Segal,E., Koller,D. and Kim,S.K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.
20. Wu,L.F., Hughes,T.R., Davierwala,A.P., Robinson,M.D., Stoughton,R. and Altschuler,S.J. (2002) Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nature Genet*., **31**, 255–265.
21. Karaoz,U., Murali,T.M., Letovsky,S., Zheng,Y., Ding,C., Cantor,C.R. and Kasif,S. (2004) Whole genome annotation using evidence integration in functional linkage networks. *Proc. Natl Acad. Sci. USA*, 2888–2893.
22. Lee,I., Date,S.V., Adai,A.T. and Marcotte,E.M. (2004) A probabilistic functional network of yeast genes. *Science*, **306**, 1555–1558.
23. Marcotte,E.M., Pellegrini,M., Thompson,M.J., Yeates,T.O. and Eisenberg,D. (1999) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.
24. Yanai,I., Mellor,J.C. and DeLisi,C. (2002) Identifying functional links between genes using conserved chromosomal proximity. *Trends Genet*., **18**, 176–179.
25. Gary,D. Bader, Betel,D. and Christopher,W.V. Hogue (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*., **31**, 248–250.
26. Hermjakob,H., Montecchi-Palazzi,L., Lewington,C., Mudali,S., Kerrien,S., Orchard,S., Vingron,M., Roechert,B., Roepstorff,P., Valencia,A. *et al*. (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res*., **32**, D452–D455.
27. McDermott,J. and Samudrala,R. (2003) Bioverse: Functional, structural and contextual annotation of proteins and proteomes. *Nucleic Acids Res*., **31**, 3736–3737.
28. Mellor,J.C., Yanai,I., Clodfelter,K.H., Mintseris,J. and DeLisi,C. (2002) Predictome: a database of putative functional links between proteins. *Nucleic Acids Res*., **30**, 306–309.
29. Ramani,A.K., Bunescu,R.C., Mooney,R.J. and Marcotte,E.M. (2005) Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol*., **6**, R40–R40.12.
30. Breitkreutz,B.-J., Stark,C. and Tyers,M. (2003) The GRID: the General Repository for Interaction Datasets. *Genome Biol*., **4**, R23.
31. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Medhedov,S.L. and Nikolskaya,A.N. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.

32. von Mering,C., Jensen,L.J., Snel,B., Hooper,S.D., Krupp,M., Foglierini,M., Jouffre,N., Huynen,M.A. and Bork,P. (2005) STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Res*., **33**, D433–D437.
33. Xenarios,I., Fernandez,E., Salwinski,L., Duan,X.J., Thompson,M.J., Marcotte,E.M. and Eisenberg,D. (2001) DIP: the database of interacting proteins: 2001 update. *Nucleic Acids Res*., **29**, 239–241.
34. Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G., Helmer-Citterich,M. and Cesareni,G. (2002) MINT: a Molecular INTeraction database. *FEBS Lett*., **513**, 135–140.
35. Fraser,A.G. and Marcotte,E.M. (2004) A probabilistic view of gene function. *Nature Genet*., **36**, 559–564.
36. Bergmann,S., Ihmels,J. and Barkai,N. (2003) Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol*., **2**, E9.
37. Hughes,T.R., Marton,M.J., Jones,A.R., Roberts,C.J., Stoughton,R., Armour,C.D., Bennett,H.A., Coffey,E., Dai,H. and He,Y.D. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.
38. Stelzl,U., Worm,U., Lalowski,M., Haenig,C., Brembeck,F.H., Goehler,H., Stroedicke,M., Zenkner,M., Schoenherr,A., Koeppen,S. *et al*. (2005) A human protein–protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–968.
39. Rual,J.F., Venkatesan,K., Hao,T., Hirozane-Kishikawa,T., Dricot,A., Li,N., Berriz,G.F., Gibbons,F.D., Dreze,M., Ayivi-Guedehoussou,N. *et al*. (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, **437**, 1173–1178.