A Quick Guide to Large Scale Genomic Data Mining

Curtis Huttenhower[1,*] and Oliver Hofmann[1]
[1]Department of Biostatistics, Harvard School of Public Health
[*]To whom correspondence should be addressed: chuttenh@hsph.harvard.edu

Introduction

For the first several hundred years of research in cellular biology, the main bottleneck to scientific progress was data collection. Our newfound data-richness, however, has shifted this bottleneck from collection to analysis [1]. While a variety of options exists for examining any one experimental dataset, we are still discovering what new biological questions can be answered by mining thousands of genomic datasets in tandem, potentially spanning different molecular activities, technological platforms, and model organisms. As an analogy, consider the difference between searching one document for a keyword and executing an online search. While the tasks are conceptually similar, they require vastly different underlying methodologies, and they have correspondingly large differences in their potentials for knowledge discovery.

Large scale genomic data mining is thus the process of using many (potentially diverse) datasets, often from public repositories, to address a specific biological question. Statistical meta-analyses are an excellent example, in which many experimental results are examined in order to lend statistical power to a hypothesis test (e.g. for differential expression) [2,3]. As the amount of available genomic data grows, however, exploratory methods allowing hypothesis generation are also becoming more prevalent. The ArrayExpress Gene Expression Atlas, for example, allows users to examine hundreds of experimental factors across thousands of independent experimental results [4]. In most cases, though, an investigator with a specific question in mind must collect relevant data to bring to bear on a question of interest. Some examples might be:

- If you've obtained a gene set of interest, in which tissues or cell lines are they coexpressed?
- If you assay a particular cellular environment, are there other experimental conditions that incur a similar genomic response?
- If you have high-specificity, low-throughput data for a few genes, with what other genes do they interact or coexpress in high-throughput data repositories? Under what experimental conditions, or in which tissues?

Bringing large quantities of genomic data to bear on such questions involves three main tasks: establishing methodology for efficiently querying large data collections; assembling data from appropriate repositories; and integrating information from a variety of experimental data types. Since the technical [5,6,7] and methodological [8,9,10] challenges in heterogeneous data integration have been discussed elsewhere, this introduction will focus mainly on the first two points. As discussed below, the computational requirements for processing thousands of whole-genome datasets in a reasonable amount of time must be addressed, either algorithmically or using cloud or distributed computing [11,12]. Subsequently, data collection is sometimes easy - as is increasingly the case for high-throughput sequencing, individual experiments can themselves be the sources of large data repositories. In other cases, a biological investigation might benefit from the inclusion of substantial external or public data.

Methods and pitfalls in manipulating genomic data

A point that must be emphasized when dealing with very large genomic data collections is that many convenient computational tools for individual dataset analysis will scale poorly to repositories of hundreds or thousands of genome-scale experimental results. Scripting environments such as R/Bioconductor [13] and MATLAB (The MathWorks, Natick, MA) should be used with caution to avoid excessive runtimes. Similarly, data storage can be as great or greater a concern as data processing: plain text or XML storage formats, while conveniently human-readable, can waste unsustainable amounts of space for large repositories.

Solutions to these technical issues include software and data access methodologies specifically tailored to large scale data manipulation. Three broad categories of solutions exist: web applications that aggregate information from multiple sources, programmatic APIs that allow sophisticated computational queries of individual large data sources, and do-it-yourself solutions that rely on manually obtaining and processing bulk data from public repositories. In the first category, most current bioinformatic systems include online interfaces, but these generally provide analyses of individual datasets rather than large compendia. Notable exceptions include the STRING [14] and BioMart [15] tools, which aggregate a large number of functional and sequence annotation data sources, respectively. Integrated results and data portals are also available for many model organisms, including HEFalMp [16], Endeavour [17], and the Prioritizer [18] for human data, integrated within- [19] and across-species [20] results for *C. elegans*, bioPIXIE [21] and SPELL [22] for *S. cerevisiae*, and a variety of tools for other systems [23,24,25].

While these online tools provide pre-computed data mining results, a second option is to perform tailored queries of experimental results from one or more large public repositories. This adds a level of complexity, since you must still decide on appropriate downstream analyses of the retrieved data, but the heavy lifting of data normalization, filtering, and search is still done by the remote system. Manual portals to such information are the core of canonical interfaces at the NCBI [26] and EBI [27], and workflow systems such as Taverna [28] and Galaxy [29] are emerging to automate significant portions of these analysis pipelines. Most major data repositories now offer programmable interfaces using one of several common protocols: HTTP (i.e. programmatic URLs or REST) [26,27], SOAP [30,31], or bioinformatic services such as DAS [7], BioMOBY [32], or Gaggle [33]. These protocols provide a way to pose sophisticated queries to a data repository, leaving you to examine only the end products of interest.

The greatest level of flexibility in large scale biological data mining is offered by manually processing bulk experimental data, which of course also incurs the greatest level of time commitment and overhead. However, this is currently one of the only ways in which sophisticated multifactorial queries can be executed. If you're interested in identifying potential targets of yeast cell cycle kinases under a variety of culture growth conditions, even a relatively complex large scale computational screen will likely be simpler than running new corresponding high-throughput assays:

1. By examining the *S. cerevisiae* GO [34] annotations at the Saccharomyces Genome Database [35], we find that the intersection between the *cell cycle* process (669 genes) and the *protein kinase activity* function (135 genes, both terms downloadable at AmiGO [36]) yields a list of 51 genes.
2. By downloading the DIP [37], MINT [38], and bioGRID [39] interaction databases (discussed below) in bulk and searching for all interactions in which these genes' products participate, we obtain 7,830 potential kinase-target pairs.

3. By downloading all GEO [40] yeast expression data in bulk (also discussed below), calculating all normalized correlations using Sleipnir ([11], a calculation taking <1hr), and listing only correlations stringently significant at a corrected 0.01 level ($p=1.2x10^{-5}$, $z=4.22$), we find 81 cell cycle kinase-target pairs with high correlation under some experimental condition.
4. It is vital to evaluate the accuracy of our predictions, although since GO was used as part of the input data, care must be taken to avoid a circular evaluation. In this case, the non-kinase interaction partners were predicted solely based on experimental interactions and coexpression, and we find that 45 of them (~25%, hypergeometric $p<10^{-8}$) indeed have known roles in the cell cycle.

Note that in each of these steps, experimental data of several different types is processed using a uniform network model, and this workflow for large scale biological data analysis is summarized in Figure 1; a description of the analysis is provided in Box 1 and detailed commands in Text S1. This small example is obviously biologically somewhat naive, but it demonstrates the remarkably nuanced questions that can be answered using large scale data mining even without complex machine learning methodology.

Unsurprisingly, a number of common technical pitfalls arise in large scale data analysis. Even structured databases can break down in the face of thousands of whole-genome interactomes, leading most current large-scale data repositories to employ some combination of filesystem-based flat file storage archives and binary formats (including GenBank's ASN.1 PER [26], BioHDF [41], and Sleipnir's DAB [11]). Data transfer mechanisms for bulk data are often limited to FTP or Aspera (http://www.asperasoft.com), although experimental metadata is often available through sophisticated programmable interfaces [40,42,43]. Several reviews have been written dealing with inter-study data normalization [8,44], particularly for microarrays [45,46,47] - although perhaps the simplest yet most important normalizations required are often chromosomal coordinates and gene, transcript, and protein identification schemes [48].

Genomic data resources

Three practical impediments to large scale integrative data mining are data availability, data size, and algorithms and models for integration. As discussed above, the challenges inherent in manipulating large data can often be overcome through compact encodings and awareness of efficiency issues. Similarly, although many sophisticated systems for biological data integration exist [8,9,10,49], they are not always necessary in order to discover new biology in large data collections. As demonstrated by the toy analysis above, simply asking the right questions of several different data repositories can rapidly generate novel biological hypotheses. It remains to discover and catalog the availability and scope of these repositories; the annual Nucleic Acids Research database issue [50] is an excellent resource for this, as are online database aggregators (e.g. [51,52,53] and http://biodatabase.org), and several primary biological data types and sources are presented here in summary.

High-throughput sequencing

Next-generation short-read DNA sequencing is rapidly becoming a current-generation technology and producing ever-longer read lengths. While the purpose of this manuscript is not to address the (serious) informatic requirements needed for processing raw sequence data, several points raised by [1] are worth summarizing. Current sequencers can generate up to 400 million 50-100bp reads per run, and this number will be obsolete soon after this manuscript is published. Performing even the simplest analyses on this data, let alone assembly, polymorphism detection, annotation, or other complex tasks,

requires sophisticated computational hardware *and* software. Few cookie-cutter solutions are available, given how rapidly the technology continues to change, but online forums such as SEQanswers (http://seqanswers.com) are currently one of the best resources for up-to-date information on short-read sequencing.

When investigating individual organisms' genomes (discussed below in more detail), many of the tools for large scale sequence mining are focused on the study of variation: across disease state tissue or pathogen samples (e.g. The Cancer Genome Atlas [54] and the Cancer Genome Project [55]), structurally or polymorphically across individuals (e.g. the 1,000 Genomes Project [56] and the Personal Genome Project [57]), or phylogenetically across species (e.g. Genome 10K [58]). Particularly for phylogeny and evolutionary relationships, a variety of tools are available online that efficiently summarize very large sequence collections; EMBOSS [59], MEGA [60], MEGAN [61], and mothur [62] are only a few of the creatively-named systems available in this area.

An interesting large scale data mining opportunity afforded by modern sequencing techniques is provided by metagenomic repositories such as CAMERA [63], MG-RAST [64], and IMG/M [65], all of which offer tools for inter-study comparisons of multiple environmental or microfloral datasets. For instance, an experimenter can easily upload an entire metagenome to MG-RAST and receive a detailed profile of the community's metabolic potential; using CAMERA, fragment recruitment profiles can be generated comparing any pair of metagenomes. Simultaneously considering the functional diversity of a metagenome, its constituent organisms, and the associated experimental metadata allows a single analysis to scale from molecular mechanisms to global ecology [66].

Whole-genome sequences

The first widely-used large scale biological data repositories were (arguably) for reads deposited during the Human Genome Project and other pioneering sequencing projects, and these remain important sources of annotated genomic sequences. GenBank [67] has diversified to include a variety of online and offline tools such as the Genome Workbench, and Ensembl [68] provides an invaluable online window onto a number of genome builds. The Sanger Institute hosts a number of additional genome resources (http://www.sanger.ac.uk/Projects/), and the JGI provides several microbial genomes and associated tools [69]. Sequence annotations have been reviewed elsewhere [70] and include everything from open reading frames through regulatory sites to chromatin structure and epigenetics; much of this information is available through a uniform interface at the UCSC Genome Browser [71]. Sequence data has been highly standardized over the years, with most raw sequences provided as FASTA or its variants, detailed annotations provided as GenBank/EMBL files, and brief annotations as GFFs. Most sequence manipulation software will recognize all of these formats [72].

Microarrays

Similarly, gene expression microarrays were the first functional data to be analyzed on a large scale, although applications of high-throughput sequencing are poised to overtake them in widespread data availability. The Gene Expression Omnibus [40] and ArrayExpress [42] databases are the most common sources of array data, with Celsius [73], field-specific resources such as Oncomine [74], and institute-specific databases [75] providing additional datasets. Both GEO and ArrayExpress provide programmatic interfaces and structured FTP filesystems for bulk analysis. GEO data is standardized around the SOFT text file format [40] and ArrayExpress around the MGED MAGE format family [76];

both are variants of tab-delimited text and can be manipulated by a variety of publicly available tools [77,78] or custom software.

Physical, genetic, and regulatory interactomes

Interactomes are significantly more diverse than sequence and expression data, both in their biological grounding and their electronic availability and distribution. For a subset of the many available physical, genetic, and regulatory interaction databases, we refer the reader to previous articles in the PLoS Computational Biology Getting Started series [79]. These data are distributed in a diversity of formats and with a variety of experimental metadata. The fundamental computational data being communicated is most often an unweighted (possibly directed) graph, and interactome data thus lends itself well to large scale exploration using simple Boolean operations and graph mining algorithms [80,81]. More biologically focused investigation can be done using, for example, PSI-formatted files containing experimental and biological metadata [82].

Other genomic data types and sources

This is only a small selection of the data resources that can be mined integratively to address biological questions, with structural [83,84], proteomic [85,86], and metabolic [87] databases being obvious large scale omissions. A final data type that must be considered, however, is not directly experimental; curated pathway and structured knowledge resources are invaluable in the planning and validation of large scale data mining [34,88,89,90]. Two vital considerations when using such resources are, first, that they are originally based on published literature and experimental results. Subtle issues of circularity can arise when curated resources are used to supplement or validate data mining results, since the data being analyzed may itself have contributed to the curation process. Second, we have as yet to discover and catalog all biological knowledge - when used as gold standards, even the best-curated resources can be incomplete in the face of the billions of datapoints now being generated by the field on a regular basis, with important consequences in computational learning and evaluation [91].

Outlook

With almost every type of biological data accumulating at an exponential rate, large scale genomic data mining is increasingly becoming a necessity. For computational investigators, this represents a clear opportunity for methodology development; since data are becoming available at a rate that outpaces even Moore's law, it is not enough to wait for faster computers to execute longer and longer queries, and new bioinformatic tools must be developed with an eye to scalability and efficiency (e.g. through massive parallization). However, the opportunity for biological investigation is at least as large. Nature has already harnessed scalability to her own advantage, and the combinatorics of the genetic code, multimodal and combinatorial regulation, cellular differentiation, and temporal development ensure that even our current wealth of data provides an incomplete view of biological complexity. A simple justification for broad-ranging computational screens of genomic data is their speed and low cost as a precursor to more extensive laboratory work. An even more compelling motivation, though, is the fact that the extent and complexity of biological systems may best be discovered by simultaneously considering a wide range of genome-scale data.

Acknowledgements

We would like to gratefully thank Winston Hide and Olga Troyanskaya for their input into this tutorial and the PLoS editorial team and reviewers for supporting the Education Collection.

References

1. McPherson JD (2009) Next-generation gap. Nat Methods 6: S2-5.
2. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, et al. (2004) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci U S A 101: 9309-9314.
3. Cahan P, Rovegno F, Mooney D, Newman JC, St Laurent G, 3rd, et al. (2007) Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization. Gene 401: 12-18.
4. Kapushesky M, Emam I, Holloway E, Kurnosov P, Zorin A, et al. (2010) Gene expression atlas at the European bioinformatics institute. Nucleic Acids Res 38: D690-698.
5. Hwang D, Smith JJ, Leslie DM, Weston AD, Rust AG, et al. (2005) A data integration methodology for systems biology: experimental verification. Proc Natl Acad Sci U S A 102: 17302-17307.
6. Butte AJ, Kohane IS (2006) Creation and implications of a phenome-genome network. Nat Biotechnol 24: 55-62.
7. Jenkinson AM, Albrecht M, Birney E, Blankenburg H, Down T, et al. (2008) Integrating biological data - the Distributed Annotation System. BMC Bioinformatics 9 Suppl 8: S3.
8. Troyanskaya OG (2005) Putting microarrays in a context: integrated analysis of diverse biological data. Brief Bioinform 6: 34-43.
9. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, et al. (2006) Gene prioritization through genomic data fusion. Nat Biotechnol 24: 537-544.
10. Lee I, Marcotte EM (2008) Integrating functional genomics data. Methods Mol Biol 453: 267-278.
11. Huttenhower C, Schroeder M, Chikina MD, Troyanskaya OG (2008) The Sleipnir library for computational functional genomics. Bioinformatics 24: 1559-1561.
12. Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL (2009) Searching for SNPs with cloud computing. Genome Biol 10: R134.
13. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80.
14. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, et al. (2009) STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37: D412-416.
15. Haider S, Ballester B, Smedley D, Zhang J, Rice P, et al. (2009) BioMart Central Portal--unified access to biological data. Nucleic Acids Res 37: W23-27.
16. Huttenhower C, Haley EM, Hibbs MA, Dumeaux V, Barrett DR, et al. (2009) Exploring the human genome with functional maps. Genome Res 19: 1093-1106.
17. Tranchevent LC, Barriot R, Yu S, Van Vooren S, Van Loo P, et al. (2008) ENDEAVOUR update: a web resource for gene prioritization in multiple species. Nucleic Acids Res 36: W377-384.
18. Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, et al. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Genet 78: 1011-1025.
19. Gunsalus KC, Ge H, Schetter AJ, Goldberg DS, Han JD, et al. (2005) Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis. Nature 436: 861-865.
20. Zhong W, Sternberg PW (2006) Genome-wide prediction of C. elegans genetic interactions. Science 311: 1481-1484.
21. Myers CL, Robson D, Wible A, Hibbs MA, Chiriac C, et al. (2005) Discovery of biological networks from diverse functional genomic data. Genome Biol 6: R114.

22. Hibbs MA, Hess DC, Myers CL, Huttenhower C, Li K, et al. (2007) Exploring the functional landscape of gene expression: directed search of large microarray compendia. Bioinformatics 23: 2692-2699.
23. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, et al. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 302: 449-453.
24. Pena-Castillo L, Tasan M, Myers CL, Lee H, Joshi T, et al. (2008) A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. Genome Biol 9 Suppl 1: S2.
25. Alexeyenko A, Sonnhammer EL (2009) Global networks of functional coupling in eukaryotes from comprehensive data integration. Genome Res 19: 1107-1116.
26. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, et al. (2009) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res.
27. McWilliam H, Valentin F, Goujon M, Li W, Narayanasamy M, et al. (2009) Web services at the European Bioinformatics Institute-2009. Nucleic Acids Res 37: W6-10.
28. Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, et al. (2006) Taverna: a tool for building and running workflows of services. Nucleic Acids Res 34: W729-732.
29. Blankenberg D, Taylor J, Schenck I, He J, Zhang Y, et al. (2007) A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly. Genome Res 17: 960-964.
30. Sand O, Thomas-Chollier M, Vervisch E, van Helden J (2008) Analyzing multiple data sets by interconnecting RSAT programs via SOAP Web services: an example with ChIP-chip data. Nat Protoc 3: 1604-1615.
31. Stockinger H, Attwood T, Chohan SN, Cote R, Cudre-Mauroux P, et al. (2008) Experience using web services for biological sequence analysis. Brief Bioinform 9: 493-505.
32. Wilkinson MD, Senger M, Kawas E, Bruskiewich R, Gouzy J, et al. (2008) Interoperability with Moby 1.0--it's better than sharing your toothbrush! Brief Bioinform 9: 220-231.
33. Shannon PT, Reiss DJ, Bonneau R, Baliga NS (2006) The Gaggle: an open-source software system for integrating bioinformatics software and data sources. BMC Bioinformatics 7: 176.
34. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.
35. Hong EL, Balakrishnan R, Dong Q, Christie KR, Park J, et al. (2008) Gene Ontology annotations at SGD: new data sources and annotation methods. Nucleic Acids Res 36: D577-581.
36. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, et al. (2009) AmiGO: online access to ontology and annotation data. Bioinformatics 25: 288-289.
37. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, et al. (2004) The Database of Interacting Proteins: 2004 update. Nucleic Acids Res 32: D449-451.
38. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, et al. (2007) MINT: the Molecular INTeraction database. Nucleic Acids Res 35: D572-574.
39. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, et al. (2006) BioGRID: a general repository for interaction datasets. Nucleic Acids Res 34: D535-539.
40. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, et al. (2009) NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res 37: D885-890.
41. Dougherty MT, Folk MJ, Zadok E, Bernstein HJ, Bernstein FC, et al. (2009) Unifying Biological Image Formats with HDF5. Bioscience 7.

42. Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, et al. (2009) ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res 37: D868-872.
43. Quackenbush J (2009) Data reporting standards: making the things we use better. Genome Med 1: 111.
44. Louie B, Mork P, Martin-Sanchez F, Halevy A, Tarczy-Hornoch P (2007) Data integration and genomic medicine. J Biomed Inform 40: 5-16.
45. Quackenbush J (2002) Microarray data normalization and transformation. Nat Genet 32 Suppl: 496-501.
46. Steinhoff C, Vingron M (2006) Normalization and quantification of differential expression in gene expression microarrays. Brief Bioinform 7: 166-177.
47. Kapushesky M, Emam I, Holloway E, Kurnosov P, Zorin A, et al. (2009) Gene Expression Atlas at the European Bioinformatics Institute. Nucleic Acids Res.
48. Durinck S, Spellman PT, Birney E, Huber W (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc 4: 1184-1191.
49. Lanckriet GR, De Bie T, Cristianini N, Jordan MI, Noble WS (2004) A statistical framework for genomic data fusion. Bioinformatics 20: 2626-2635.
50. Cochrane GR, Galperin MY (2010) The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources. Nucleic Acids Res 38: D1-4.
51. Babu PA, Udyama J, Kumar RK, Boddepalli R, Mangala DS, et al. (2007) DoD2007: 1082 molecular biology databases. Bioinformation 2: 64-67.
52. Chen YB, Chattopadhyay A, Bergen P, Gadd C, Tannery N (2007) The Online Bioinformatics Resources Collection at the University of Pittsburgh Health Sciences Library System--a one-stop gateway to online bioinformatics databases and software tools. Nucleic Acids Res 35: D780-785.
53. Brazas MD, Yamada JT, Ouellette BF (2009) Evolution in bioinformatic resources: 2009 update on the Bioinformatics Links Directory. Nucleic Acids Res 37: W3-5.
54. Network TCGAR (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455: 1061-1068.
55. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, et al. (2007) Patterns of somatic mutation in human cancer genomes. Nature 446: 153-158.
56. Hayden EC (2008) International genome project launched. Nature 451: 378-379.
57. Church GM (2005) The personal genome project. Mol Syst Biol 1: 2005 0030.
58. Scientists GKCo (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. J Hered 100: 659-674.
59. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16: 276-277.
60. Kumar S, Nei M, Dudley J, Tamura K (2008) MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform 9: 299-306.
61. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Res 17: 377-386.
62. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, et al. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75: 7537-7541.
63. Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M (2007) CAMERA: a community resource for metagenomics. PLoS Biol 5: e75.

64. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9: 386.
65. Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, et al. (2008) IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res 36: D534-538.
66. Gianoulis TA, Raes J, Patel PV, Bjornson R, Korbel JO, et al. (2009) Quantifying environmental adaptation of metabolic pathways in metagenomics. Proc Natl Acad Sci U S A 106: 1374-1379.
67. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2009) GenBank. Nucleic Acids Res 37: D26-31.
68. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, et al. (2009) Ensembl 2009. Nucleic Acids Res 37: D690-697.
69. Liolios K, Chen IM, Mavromatis K, Tavernarakis N, Hugenholtz P, et al. (2009) The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res.
70. Brent MR (2008) Steady progress and recent breakthroughs in the accuracy of automated genome annotation. Nat Rev Genet 9: 62-73.
71. Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, et al. (2009) The UCSC genome browser database: update 2010. Nucleic Acids Res.
72. Information NCfB (2009) The NCBI handbook. Bethesda, MD: National Library of Medicine.
73. Day A, Carlson MR, Dong J, O'Connor BD, Nelson SF (2007) Celsius: a community resource for Affymetrix microarray data. Genome Biol 8: R112.
74. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, et al. (2007) Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. Neoplasia 9: 166-180.
75. Demeter J, Beauheim C, Gollub J, Hernandez-Boussard T, Jin H, et al. (2007) The Stanford Microarray Database: implementation of new analysis tools and open source release of software. Nucleic Acids Res 35: D766-770.
76. Rayner TF, Rocca-Serra P, Spellman PT, Causton HC, Farne A, et al. (2006) A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. BMC Bioinformatics 7: 489.
77. Davis S, Meltzer PS (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics 23: 1846-1847.
78. Rayner TF, Rezwan FI, Lukk M, Bradley XZ, Farne A, et al. (2009) MAGETabulator, a suite of tools to support the microarray data format MAGE-TAB. Bioinformatics 25: 279-280.
79. Viswanathan GA, Seto J, Patil S, Nudelman G, Sealfon SC (2008) Getting started in biological pathway construction and analysis. PLoS Comput Biol 4: e16.
80. Huber W, Carey VJ, Long L, Falcon S, Gentleman R (2007) Graphs in molecular biology. BMC Bioinformatics 8 Suppl 6: S8.
81. Ma'ayan A (2008) Network integration and graph analysis in mammalian molecular systems biology. IET Syst Biol 2: 206-221.
82. Martens L, Orchard S, Apweiler R, Hermjakob H (2007) Human Proteome Organization Proteomics Standards Initiative: data standardization, a view on developments and policy. Mol Cell Proteomics 6: 1666-1667.
83. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, et al. (2008) Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 36: D419-425.

84. Henrick K, Feng Z, Bluhm WF, Dimitropoulos D, Doreleijers JF, et al. (2008) Remediation of the protein data bank archive. Nucleic Acids Res 36: D426-433.
85. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, et al. (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 31: 3784-3788.
86. Consortium TU (2009) The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res.
87. Oberhardt MA, Palsson BO, Papin JA (2009) Applications of genome-scale metabolic reconstructions. Mol Syst Biol 5: 320.
88. Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, et al. (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic Acids Res 33: 6083-6089.
89. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2009) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res.
90. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, et al. (2009) Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res 37: D619-622.
91. Huttenhower C, Hibbs MA, Myers CL, Caudy AA, Hess DC, et al. (2009) The impact of incomplete knowledge on evaluation: an experimental benchmark for protein function prediction. Bioinformatics 25: 2404-2410.

Figure Captions

Figure 1: Large scale genomic data mining. A schematic overview of possible inputs, data sources, network models, and output predictions from computational screens leveraging many genome-scale datasets. Note that both the "output" pathway model and the "input" experimental data are represented as networks: directed regulatory binding site targets, undirected weighted coexpression, and undirected interactions, respectively. As demonstrated by the sample analysis in Box 1, biological networks provide a uniform framework within which both experimental data and predicted models can be represented, facilitating integrative analyses.

Boxes

Box 1: An example using multiple genome-scale data repositories to determine potential kinase-target interactions active during the *S. cerevisiae* cell cycle. For step-by-step instructions on performing each task, please see Text S1.

1. Retrieve lists of known yeast cell cycle and protein kinase genes from the Gene Ontology [34] using the AmiGO [36] web service.
2. Intersect these two gene sets to find protein kinases potentially involved in the cell cycle.
3. Retrieve lists of experimentally determined protein-protein interactions from the DIP [37], MINT [38], and bioGRID [39] databases.
4. Map all appropriate gene identifiers to gene symbols using information from BioMart [15].
5. Taking the union of these three databases, identify any pairs of interacting proteins in which at least one partner is a members of the cell cycle protein kinase list. Note that this will provide a conservative underestimate, since many transient kinase-target interactions are difficult to detect based on high-throughput data.
6. Retrieve yeast expression data from GEO [40] and convert each dataset into a normalized coexpression network using the Sleipnir software [11].
7. Extract all gene pairs correlated above a multiple hypothesis corrected 0.01 significance level, and intersect these pairs with the list of cell cycle protein kinase interactions.
8. This produces a list of potential cell cycle-linked phosphorylation targets, based on protein kinases known to be involved in the cell cycle, interacting with the putative target, and coexpressing strongly with it under some experimental condition.
9. Finally, evaluate the proposed list's plausibility by examining how many of the non-kinase partners are known cell cycle genes.

Sequence Data

Gene annotations, high-throughput seq., regulatory sites...

GenBank/Ensembl/etc.

Microarray Data

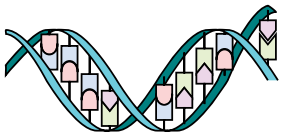
Coexpression, differential expression, CGH, SNPs, ChIP-chip...

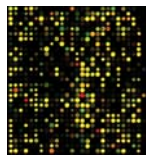
GEO/ArrayExpress/etc.

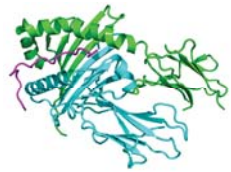
Interaction Data

Physical, regulatory, genetic, protein modifications...

BioGRID/IntAct/etc.



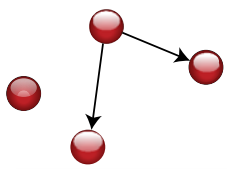


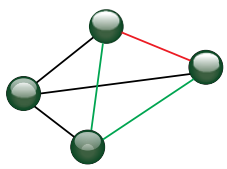


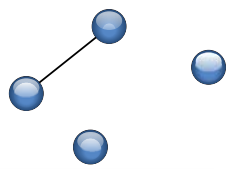
Curated Data

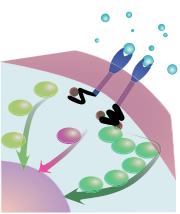
Detailed mechanistic descriptions of pathways and function

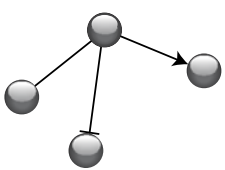
GO/KEGG/etc.





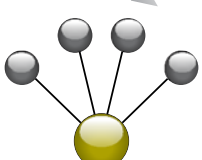


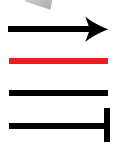


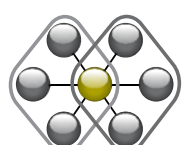


Large Scale Genomic Data Mining

Computational screens using efficient algorithms and many (potentially diverse) genome-scale datasets to generate specific biological hypotheses:







Protein Characterization

Guilt-by-assocation biochemical and functional roles

Interaction Characterization

Predicted physical/genetic/ etc. interactions

Dataset Characterization

Find similar datasets, experimental conditions

Pathway Characterization

Coordinated activity and regulatory hubs