Predictive Networks: A Flexible, Open Source, Web Application for Integration and Analysis of Human Gene Networks

D866–D875 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 16 November 2011*
doi:10.1093/nar/gkr1050

Predictive networks: a flexible, open source, web application for integration and analysis of human gene networks

Benjamin Haibe-Kains[1,2,*], Catharina Olsen[3], Amira Djebbari[4], Gianluca Bontempi[3], Mick Correll[5], Christopher Bouton[6] and John Quackenbush[1,2,5,*]

[1]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, [2]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02215, USA, [3]Machine Learning Group, Université Libre de Bruxelles, Brussels, Belgium, [4]Ontario Cancer Institute, Princess Margaret Hospital/UHN, and the Campbell Family Institute for Cancer Research, University of Toronto, Toronto, ON M5G 1L7, Canada, [5]Center for Cancer Computational Biology, Dana-Farber Cancer Institute, Boston, MA 02215 and [6]Entagen, Newburyport, MA 01950, USA

Received August 15, 2011; Revised October 9, 2011; Accepted October 23, 2011

ABSTRACT

Genomics provided us with an unprecedented quantity of data on the genes that are activated or repressed in a wide range of phenotypes. We have increasingly come to recognize that defining the networks and pathways underlying these phenotypes requires both the integration of multiple data types and the development of advanced computational methods to infer relationships between the genes and to estimate the predictive power of the networks through which they interact. To address these issues we have developed *Predictive Networks* (PN), a flexible, open-source, web-based application and data services framework that enables the integration, navigation, visualization and analysis of gene interaction networks. The primary goal of PN is to allow biomedical researchers to evaluate experimentally derived gene lists in the context of large-scale gene interaction networks. The PN analytical pipeline involves two key steps. The first is the collection of a comprehensive set of known gene interactions derived from a variety of publicly available sources. The second is to use these 'known' interactions together with gene expression data to infer robust gene networks. The PN web application is accessible from http://predictivenetworks.org. The PN code base is freely available at https://sourceforge.net/projects/predictivenets/.

INTRODUCTION

The sequencing of the human genome and the development of new approaches including genomics (DNA), transcriptomics (RNA), methylomics (epigenetic methylation) and proteomics (protein), have given scientists the tools necessary to amass comprehensive datasets of genomic profiles in a range of cellular and organismal phenotypes and in response to a variety of perturbations. While the hope was that we could use these data to understand the link between genotype and phenotype, we have increasingly come to recognize that the cellular regulatory processes are more complex than we had once imagined. We now understand that it is generally not individual genes, but networks of interacting genes and gene products, which collectively interact to define phenotypes and the alterations that occur in the development of disease.

Network models were first applied to gene expression data from a yeast cell cycle experiment in which synchronized cells were profiled over a carefully planned time-course (1). Friedman *et al*. (2) analyzed these data in a Bayesian Network framework to develop a predictive cell-cycle model. Since this early work, there have been many other methods developed to model networks while addressing the intrinsic complexity of high-throughput genomic data (high feature-to-sample ratio, high-level of noise and co-linearity) (1–9). Other web-based tools, such as ASIAN (10), SEBINI (11) and CARRIE (12), attempt to infer interaction networks based solely on genomic data. However, few methods have come into widespread use and often fail to produce useful network models (13,14). The problem may be that most methods deal

*To whom correspondence should be addressed. Email: bhaibeka@jimmy.harvard.edu
Correspondence may also be addressed to John Quackenbush. Tel: + 1 617 632 3012; Fax: +1 617 632 2444; Email: johnq@jimmy.harvard.edu

solely with genomic data and ignore what may be the best resource we have to efficiently constraint the fitting of network models: the collection of existing prior knowledge captured in published biomedical literature and structured databases.

There are a number of web-based tools have been developed to retrieve putative gene–gene interactions based on descriptions in PubMed abstracts and in biological databases, including GeneMANIA (15) and iHOP/GIM (16,17). Commercial tools, such as GeneGO (18) and Ingenuity Pathway Analysis (19), combine this functionality with enrichment analysis that allows users to estimate significance of key biological functions and processes represented among a list of genes. But these tools generally treat the networks inferred from prior knowledge as scaffolds onto which gene expression data is projected, rather than as a tool to help guide network inference using genomic data. And further, we must recognize that a 'known network' based on published information may not represent the true biological network or may fail to capture the network alterations that may be associated with the phenotypes or conditions being analyzed in a particular study.

Here we present *Predictive Networks* (PN; http://predictivenetworks.org), a flexible, open-source, web-based application and data services framework for inferring networks using gene expression data in combination with gene–gene interactions mined from the full-text biomedical literature and publicly available network and pathway databases. PN allows users to create 'phenomenological' models based on the observed data that facilitate hypothesis generation and that can help identify the most relevant genes for distinguishing between phenotypes in an analysis.

COLLECTING, INTEGRATING AND ANALYZING GENE INTERACTIONS

Gene–gene interactions are described through the action of one gene on another. For example, we can define an interaction through the sentence 'PGC is inhibited by SIRT1' or 'CCNT1 regulates PGC' which have the basic English language structure:

[*Subject*; *Predicate*; *Object*].

This structure, called a 'triple', is the basis of data representation in Semantic Web technologies (20) and as such they provide a natural way of describing and characterizing interactions and networks. For example, if we know that PGC is inhibited by SIRT1 and that CCNT1 regulates PGC then we have a simple network suggesting both SIRT1 and CCNT1 influence PGC (Figure 1A); this network is actually composed of the two triples ['PGC', 'is inhibited by', 'SIRT1'] and ['CCNT1', 'regulates', 'PGC'] extracted from a PubMed abstract or a full-text article (Figure 1B). Because there are now a large number of tools for storing, searching and manipulating triples, we chose to use triples as the basic representation of gene–gene interactions in PN.

In representing interactions, the Subject and Object are both genes. One of the challenges we faced in creating a store of interactions was the frequent use of synonyms and 'common names' for genes rather than the standard Gene Symbols. The Predicates capture the interactions between genes and include terms like 'regulates' or 'is inhibited by' that capture directional interactions (*Subject* → *Object* or *Subject* ← *Object*, respectively).

One subtle challenge in capturing and representing triples is negation, and for our triples we include a flag representing the evidence in a sentence supporting the presence or absence of an interaction, as inferred from the predicate. For example, we want to distinguish between the positive interaction, 'regulates', and the negative interaction 'does not regulate'. We also recognize that the literature often contains contradictory evidence for the direction of the interaction between two genes; one way to decide likely directionality in an interaction is to weight interactions based on the number of times they have been described in various publications.

To create the database underlying the PN web application, we used two main sources from which we extracted triples (Table 1): the published biomedical literature including both abstracts catalogued in PubMed and full-text articles available through PubMed Central, and structured biological databases including Pathway Commons (21) and functional interactions identified in the recent publication of Stein *et al.* (22). The method we used to extract triples depends on the data source. Extraction of triples from structured pathway databases requires minimal processing and conversion to capture the interactions. However, deducing triples and their context from the biomedical literature is a challenging task that required we develop a text mining pipeline (Figure 2) that combined custom parsing and text mining systems (available through sourceforge, https://sourceforge.net/projects/predictivenets) together with the LingPipe text processing library (23).

The text mining pipeline starts by extracting complete text from the PubMed Open Access/Abstracts text XML files, which are available in compressed form at ftp.ncbi.nlm.nih.gov/pub/pmc/articles.*.tar.gz. A predicate ontology is used to detect predicates in each sentence from each article. A gene name detection approach is then used to identify genes appearing on either side of the predicate. If a gene–gene interaction triple is identified a final scan is performed upon the text to detect context.

To derive contexts for gene interaction triples, we used terms and phrases used in MeSH (24) and the Gene Ontology project (25) to provide an ontological representation of context keywords. We assume that the source from which triples are being extracted provides an overarching context for those gene interactions. Given this assumption we pull contextual keywords from the entire body of each source and apply those contexts to each interaction derived from the source. In assessing the relative value of derivation of triples from full-text sources versus abstracts alone, we observed that the number of contexts per triples extracted from full-text articles is significantly larger than from abstracts (Fisher's exact test $P < 0.001$; Table 1).

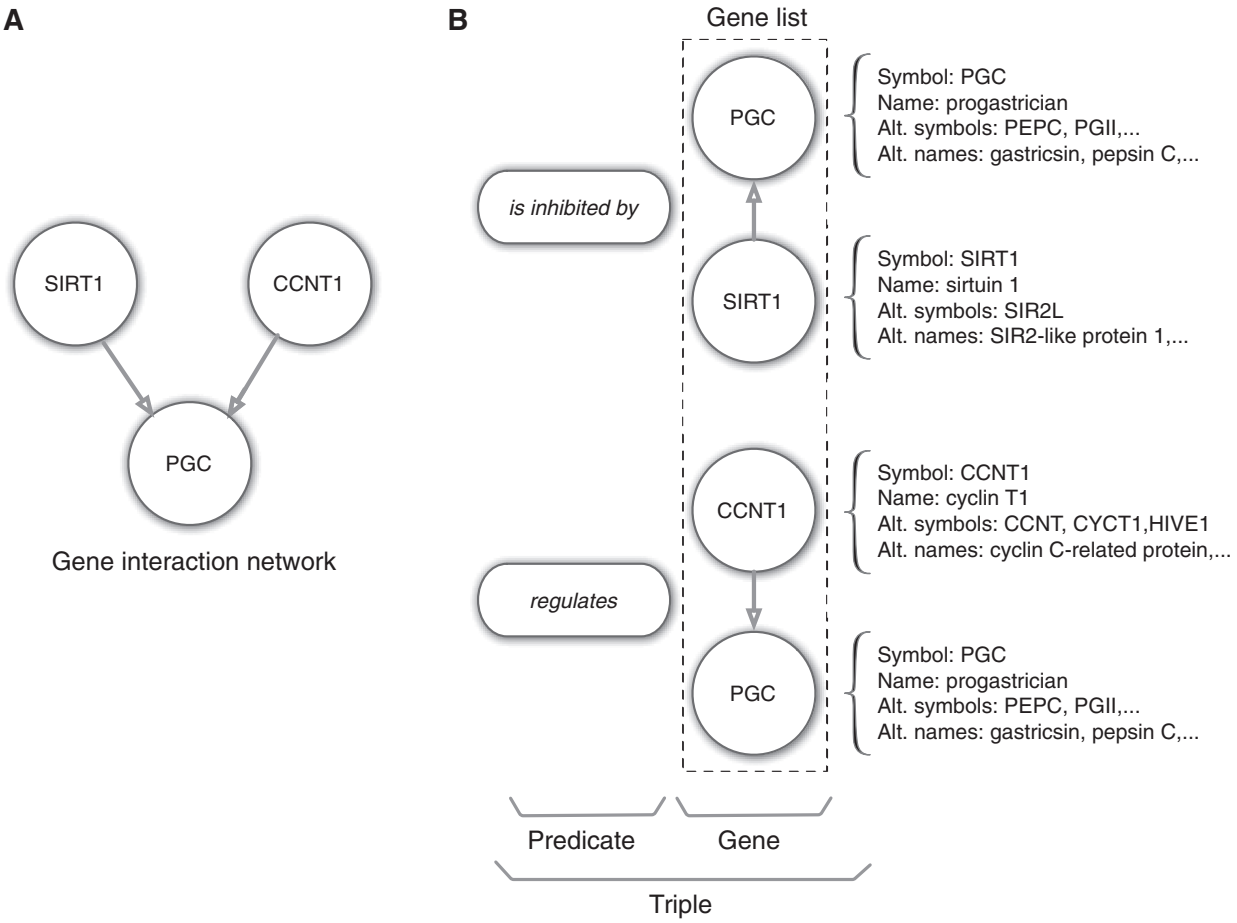

Figure 1. Overview of the core PN concepts ultimately representing a gene interaction network. A gene interaction network (**A**) is a collection of triples (**B**) where each triple involves two genes (for example PGC, SIRT1) and a predicate (for example 'is inhibited by'); each gene is described by a number of meta-data, incuding annotations; each gene can be part of a users' gene list.

Table 1. Description of data sources

	Biomedical literature		Structured biological databases	
	PubMed abstracts	PubMed Central full-text articles	Function interaction	Pathways common
No. of genes	11 443	7128	9408	11 535
No. of interactions	59 159	21 863	1 81 013	1 142 763
No. of contexts	102 987	84 507	N/A	N/A
Average no. of contexts per article	1.78	7.34	N/A	N/A
Average no. of contexts per sentence	1.53	3.59	N/A	N/A
Source	pubmed.org	pubmed.org	string-db.org	Pathwaycommons.org

The data sources used in PN including the number of genes identified, the interactions found, number of contexts, average number of such contexts per article and per sentence, and the URL from which the primary data were downloaded. No contexts have been extracted from the structured biological databases (NA, not applicable).

Use Case 1: gene-centered network searches

One of the primary questions that users of a resource such as this want to ask is 'What other genes interact with my gene of interest?' This is a relatively easy question to answer through the PN web application as all interactions are stored as triples and a network can be represented as a collection of triples connected either through common Subjects or Objects (Figure 1A and B).

The PN web application allows users to enter single gene names and to find genes that connect to it, and then genes that connect to its connection partners up to whatever distance one might desire. The PN web application allows users to submit single gene names as queries. These are disambiguated and mapped to standard identifiers which are then queried against the collected PN triples. The resulting network built from interactions

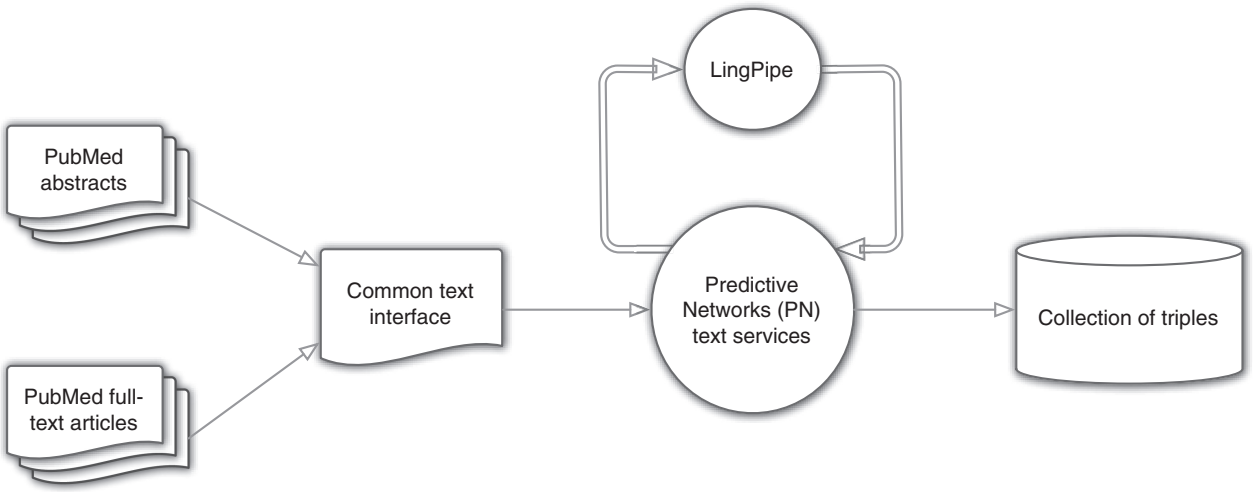

Figure 2. Text mining pipeline. PubMed abstracts and full-text articles from PubMed Central are extracted and formatted so they can be analyzed using the common text interface. The processed text is then mined using a combination of custom parsing scripts and the LingPipe text processing library to identify gene interaction triples and their contexts. These triples are then used to infer gene interactions networks used throughout PN.

represented in the triples is then visualized as a graph that can then be filtered based on the following criteria:

- Context: Networks can be limited to interactions reported in a specific context. For example a network centered on gene 'RB1' could be built using only triples identified in the context of 'tumor suppressor'.
- Number of occurrences: Networks can be built only using interactions that are supported by a sufficient number of triples. For example, users can specify that only interactions reported more than five times should be used to build the network.
- Connectivity: The default network that is presented only contains the gene and its primary (first degree) connections. Users can increase the connectivity to display more distant inferred connections to their target gene of interest and
- Max Genes: Networks can be limited to display a maximum number of genes; genes involved in the interaction supported by the largest number of triples will be selected. The default maximum of genes is 20.

In the network display, users can click on a gene name to highlight that gene and its connections. Clicking on a gene node in the network expands the network to include neighbors of that gene. Networks can be exported as a table (comma-separated value or tab-delimited spreadsheets), an image or a GML file [Graph Modeling Language (26)] which can be imported into Cytoscape (27,28) or NAVIGaTOR (29) to allow further network exploration.

Use Case 2: identifying connections in gene lists

Most genomic analyses compare two or more phenotypic conditions and report a list of genes that are significant for distinguishing between experimental groups, but the connections between the genes in the 'gene list' are often unknown. The PN web application allows users to upload a gene list and, based on user-defined parameters, searches for reported interactions stored as triples that link the genes in the list. The resulting gene interaction network is displayed in a graphical format with similar features to that described for single genes and can be exported in a variety of formats for presentation or further exploration.

Use Case 3: network inference from literature-inferred interactions and genomic data

Inferring networks from genomic data is a challenging problem. Learning the network architecture without some prior constraints requires consideration of all possible pairwise connections between genes, a problem that has been shown to be NP-hard (30). Indeed, for directed networks, the number of possible graphs is super-exponential in the number of nodes; for eight genes, there are 7.8×10^{11} network configurations that must be considered to completely explore the problem's 'state space'. Consequently, most approaches are prone to overfitting and fail to find 'realistic' networks.

Our primary goal in developing the PN application was to analyze genomic data using literature inferred relationships to help build predictive network models. Consequently, the PN application allows users to provide their own gene expression data and to use one of two methods to deduce network interactions. We previously described a seeded Bayesian network approach and demonstrated that the use of prior information could significantly improve the quality of the inferred network models (9). But inferring Bayesian networks for large numbers of genes, even with a seeded prior remains computationally challenging. Regression-based methods are not able to represent the full joint distribution of the data as Bayesian networks do, but they have been shown to have promising performance and to enable inference of extremely large biological networks (31).

To partially address the problem of network inference, the PN web application includes the seeded Bayesian network inference we had described previously (9) and a novel regression-based technique (32).

Both Bayesian and regression network inference starts with a normalized gene expression dataset and a 'Gene List' selected by users based on their own criteria. The gene expression data might be generated with any microarray platform and normalization techniques, although the use of Affymetrix (http://www.affymetrix.com) data normalized by (f)RMA (33,34) is a good choice given that they are widely used. The gene list might be a set of genes determined to be statistically significant for distinguishing sample groups or the genes in a pathway they would like to explore in the context of their expression data. Both methods use the gene list to deduce an initial network based solely on the PN triples as a starting point (or set of 'priors') that can evolve based on relationships represented in the users' gene expression data, allowing new interactions between genes to be identified or poorly supported interactions to be eliminated. Minimum and maximum size for gene expression datasets and gene lists are further described in the Supplementary Data section.

For both methods, users upload their gene expression data and gene lists, and provide a name for their analysis. The regression-based method is run by default, but users can select 'Advanced Options' to select a specific method and to set parameters for the analysis; a detailed description of the parameters is provided in the Supplementary Data section. The inferred interactions are visualized and the networks and associated data can be downloaded for further analysis.

Because the quality of inference is not uniform over the network, we developed both interaction-specific and gene-specific statistical measures to assess the quality of the constituent subnetworks that comprise the full network, This allows users to focus on the subnetworks, interactions and genes that are well supported by the data. A description of these statistics is provided in the Supplementary Data section.

The PN analytical tools rely on two interaction-specific measures, the interaction relevance and interaction stability statistics that identify gene–gene interactions that can be inferred from the prior and experimental data with high confidence. The interaction relevance statistic spans [0,1] where values close to 1 represent interactions that are strongly supported by the data (literature-inferred interactions and/or genomic data) while values close to 0 denote interactions that are only weakly supported. The interaction stability statistic also spans [0,1] and uses k-fold cross validation to estimate how strongly the inference of an interaction depends on the initial dataset used to identify it. Values close to 1 represent stable interactions that are inferred irrespective of the precise dataset used to identify them, and values close to 0 represent interactions that are rarely identified and therefore depend strongly on the dataset.

We also implemented two gene-specific statistics, R^2 [R squared (35)] and the MCC [Matthews Correlation Coefficient (36)], that both help to identify genes in the network whose expression can be reliably predicted based on the expression of their parents. The R^2 prediction score spans [0,1] where values close to 1 represent strong relationship between parent genes and a target gene as estimated by a linear regression model while values close to 0 denote weak association. R^2 is available for regression-based network inference only since this technique does not require discretization of the gene expression values (32). The MCC prediction score also spans [0,1] and its interpretation is the same than R^2 but in a classification framework where expression values are discretized. MCC is particularly suited to assess quality of Bayesian network inference because, in their traditional implementation, Bayesian network inference requires expression data to be discretized.

Use Case 4: developing a collaborative framework for defining networks

We realized that users might want to share information with colleagues about the networks they are deducing through their analysis. In the PN web application, we have implemented a community-based tool to allow collaborative development of a comprehensive gene interaction network knowledge base. While a considerable amount of information can be derived from structured data sources and text mining analyses, we recognize that expert curation can provide greater data depth and accuracy than automated methods can provide. To this end, we allow users to upload network data sets into the system and to 'publish' these so that others can use them.

WEB INTERFACE

Architecture of web interface to PN

The web interface to PN (http://predictivenetworks.org) uses the Groovy/Grails web application framework (37,38) and Java 1.6 technologies. The application runs on an Apache Tomcat 6 web application server on a CentOS 5 Linux server. Front-end interactivity makes use of the jQuery 1.3.2 Javascript library. MySQL is used as the backend database system.

PN is intended to be used for network inference both via the web application user interface and through third-party applications. In addition to the web application, the PN system also includes a RESTful API. For example, a gene symbol query to this API will return a list of all gene interactions involving that gene allowing users to import these PN triples into their own applications.

Querying PN

There are a number of starting points on the main PN page (Figure 3). Users can search for a single gene, the interaction between a pair of genes, interactions linking genes in a list and a network inference analysis panel that allows users to upload their gene expression data and a gene list to infer a robust gene interaction network. Users also have the option of creating an account and logging in so that they have access to additional

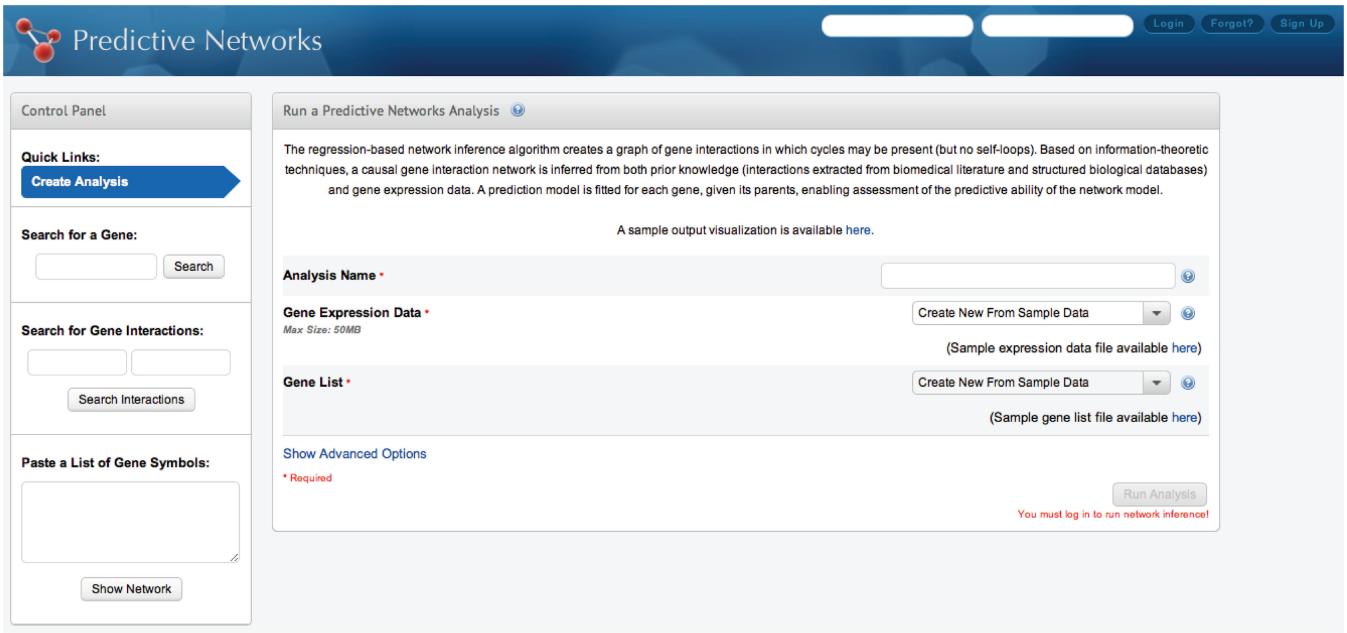

Figure 3. Front page of the PN web application displaying the four entry points: the single gene, single gene–gene interaction and gene list searches, and the network inference analysis panel. The top left panel provides a series of quick links to ensure easy navigation between the different web pages which compose the PN web application.

information as well as the option to store and revisit their analytical results and gene sets through '*My Page*'.

Network visualization

All network visualization is handled through the same basic interface, with a few additional for inferred networks. The primary display presents interactions between genes as a directed graph and provides options to filter interactions so that large graphs remain readable and avoid the 'hairball' effect (39) while remaining more relevant to the problem being studied. Among these options, users can highlight a specific gene and its interactors, select the maximum of number of genes to display (the genes involved in the largest number of triples are selected), the connection degree, filter the interactions by the number of triples that support them, directionality to a target gene (single gene search only) and the contexts in which triple have been identified.

Collaboration

Registered users have access to *My Page* through which they are able to upload gene lists, gene expression data as well as their own gene interaction networks to the system. Gene lists and networks can be kept private or shared with a defined group of users.

Documentation

All options are thoroughly documented through tooltips that provide context-specific information and options when the user 'mouses over' a relevant item. A video tutorial is available directly from the web application and a PDF manual is being prepared describing the site and its use in greater detail.

Downloading networks

All data, except users' private data, are available for download through the page *System Info*. The gene–gene interactions (triples) identified by PN are available as flat files (resource description framework, RDF files). Lists of genes can be downloaded as spreadsheets (comma separated value, CSV, or tab-delimited text files). Network visualization and heatmaps representing network properties and statistics can be exported as portable network graphics (PNG) files. Networks can also be exported as GML files for further analysis.

Database core attributes

The database core attributes are listed in Supplementary File S1, in accordance with the BioDBcore standards (40,41).

CASE STUDIES

Case Study 1: search for genes known to interact with a specific gene

Let's consider that a user wants to retrieve all the genes known to interact with BRCA1, a gene whose mutation is well-established to significantly increase risk of breast cancer (42). On the front page of the PN web application in the 'Search for a Gene:' box, the user can enter the gene symbol, 'BRCA1'. Since a gene symbol may be ambiguous, if only a gene symbol is being entered a list of

alternative gene symbols is proposed while the user is typing.

Hitting the 'Search' button returns a list of possible hits, highlighting gene names. In this case, the user can click on 'BRCA1' which brings up the option to 'View Details'. Selecting this returns detailed information on the gene, including its gene symbol, name, alternative symbol, alternative names and External IDs including EnsEMBL and EntrezGene identifiers. A first degree directed network presenting the genes most commonly reported to interact with BRCA1 is also presented, as is a gene name 'tag cloud' and a comprehensive list reporting the triples, their number of occurrences, their context and the primary data source from which they were derived. As noted previously, the network graph can be saved or exported for further analysis. All supporting documentation, including the complete list of triples containing BRCA1 can be downloaded in a variety of formats.

For BRCA1, PN identified 856 triples, of which 208 were extracted from PubMed abstract and full-text articles and the remainder were annotated in pathway databases. Three genes—BARD1, P53 and RAD51—were reported to interact with BRCA1 in at least 10 triples.

Case Study 2: network analysis of a published gene signature

There have been thousands of published studies that compare gene expression patterns between various phenotypic conditions. The product of the analysis of the data in these studies is typically a list of genes that the authors of the study found to be significant in distinguishing their experimental groups. For example, we recently discovered a gene signature that is predictive of the mutational status of PIK3CA oncogene in breast cancer (43). Although the signature is robust and highly predictive, little is known about relationships between the genes themselves and how they interact in the mutated and wild-type PIK3CA phenotypes.

We used the PN web application to infer a robust gene interaction network among these genes using both known gene–gene interactions and the gene expression data from Loi *et al.* (43). We first created a user account (only username and email address are required), logged in and went to *My Page*. We started with the 'Create Analysis' option and entered the PIK3CA gene list, which we had downloaded from GeneSigDB (http://compbio.dfci.harvard.edu/genesigdb/signaturedetail.jsp?signatureId=20479250-TableS4) (44), and uploaded a data file that included gene expression data on those 30 genes in the PIK3CA signature measured in 148 breast cancer patients (the top 30 genes of the PIK3CA signature and the gene expression data have been preloaded in the PN web application as examples) (43,45). Hitting the 'Run Analysis' button launches a network inference, the results of which are then posted to the 'Results For My Analyses' section of 'My Page'.

Clicking on the name of the analysis brings up a summary and we can further explore the results of the analysis. The topology of the inferred network, and its related statistics can be visualized through directed graphs and heatmaps (Figure 4). The page displaying

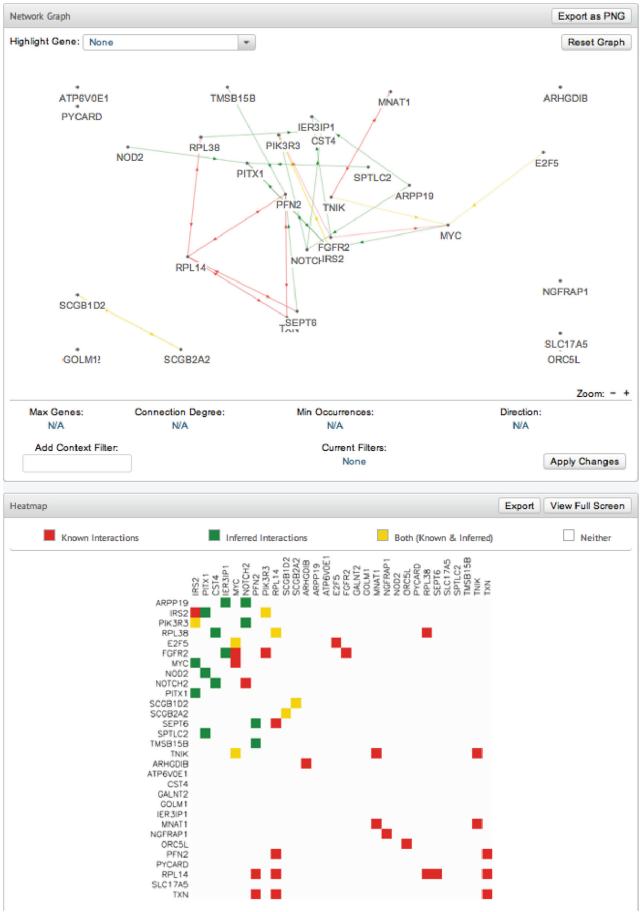

Figure 4. Gene interaction network inferred from breast cancer gene expression data and the PIK3CA signature. The network graph in the upper panel allows users to view the topology of the inferred gene interaction network. Each gene–gene interaction is color-coded to represent the evidence supporting it: literature-inferred interactions that are not supported by reported PN triples are colored in red, those inferred from the data only are green, and those supported by both are yellow. In the lower panel a color-coded heatmap allows users to quickly identify clusters of interactions and click on an interaction of interest to highlight it on the network graph.

the topology of the networks helps the user easily identify the interactions that are supported both by the priors (biomedical literature or structured biological databases or both) and the gene expression data, as well as novel interactions deduced in the analysis (Figure 4).

For the analysis described here, we find evidence of interactions for >20 of the initial 30 genes, with genes like PITX1, IRS2 or NOTCH2 being highly connected. The user can also view representations of the interaction- and gene-specific statistics and so choose to focus on the parts of the network that are inferred with higher confidence. One can see that, due to the small sample size, many interactions are unstable (stability < 0.5) but that there are also a set of highly stable interactions such as IRS2→PIK3R3 and SCGB2↔A2SCGB1D2 (Figure 5A). By looking at the R^2 prediction statistics one can see that a third of the genes have good prediction scores, which means that, despite the small sample size, the inferred

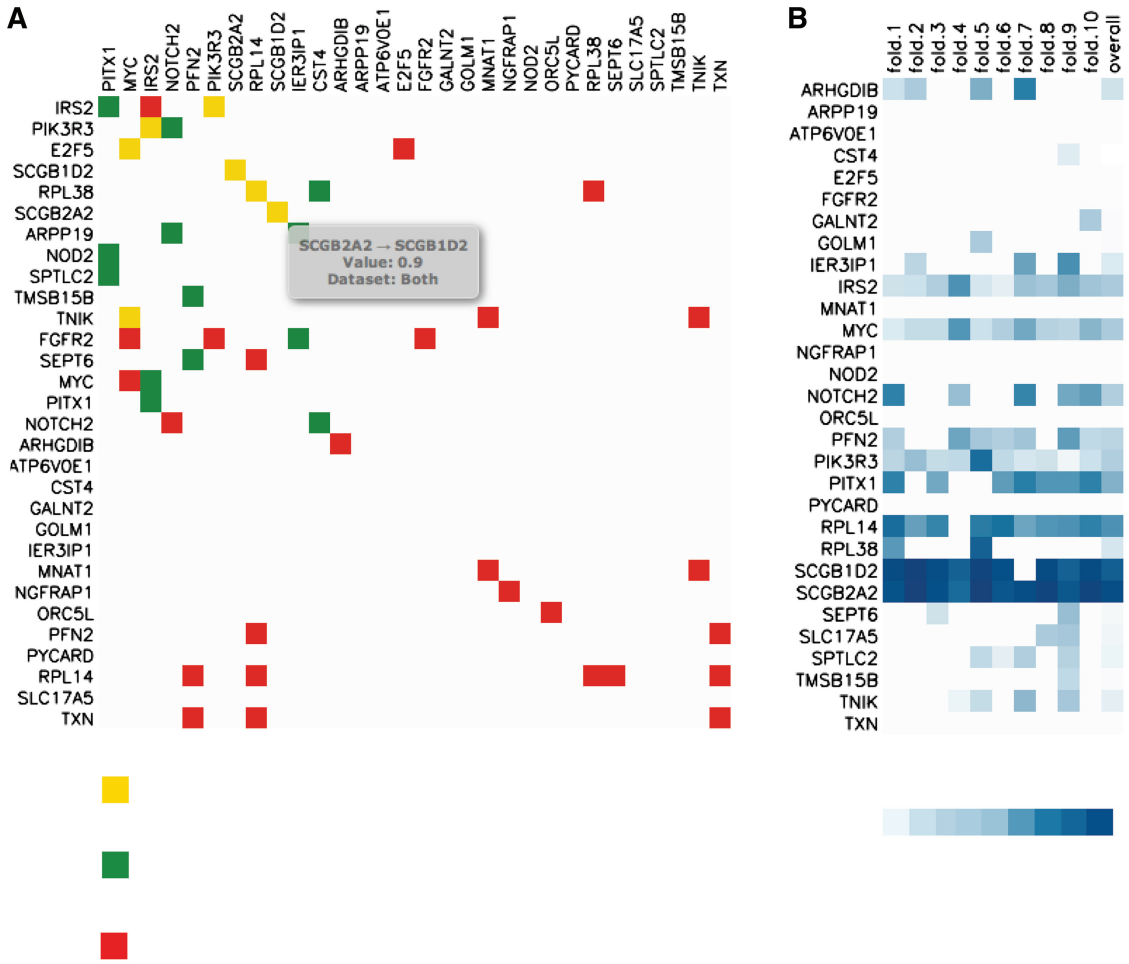

Figure 5. Interaction- and gene-specific statistics for the PIK3CA gene interaction network. The interaction-specific stability scores are represented in (**A**) and can be displayed in the PN application by mousing over the heatmap. One can see that many interactions are unstable (stability < 0.5) meaning that they cannot be confidently inferred from the data. However, some such as SCGB2A2SCGB1D2 are highly stable given the data. The gene-specific R^2 prediction scores are displayed in (**B**) where one can see that some genes can be accurately predicted given their parents, see SCGB2A2, SCGB1D2, RPL14, PITX1, IRS2, MYC and NOTCH2 for instance.

network model can reasonably predict the expression of these genes based on their parents (Figure 5B).

SUMMARY AND FUTURE DIRECTIONS

The Predictive Networks application fills an important need by allowing efficient extraction of contextualized gene–gene interactions from a variety of data sources, including full-text PubMed articles and structured biological databases. It further enables the use of these relationships to infer robust gene interaction networks from high-throughput genomic data. We have implemented these functions in an open-source web application that facilitates query and extraction of such biological networks; the framework we developed could be easily extended by power users to include additional network inference algorithms. Further, PN provides a collaboration framework in which scientists can search, share and curate gene networks.

The PN web interface will continue to develop and we intend to implement several new features, including allowing users to curate triples identified by the PN text-mining pipeline and share them with other users. We also plan to enrich the network visualization with new layouts (hierarchical or circular network representations, for instance), and to optimize visualization of heatmaps to better assess the high quality subtnetworks within very large networks. Parallelization of our network inference algorithms is also a priority as it will accelerate these computationally intensive analyses. Finally, our hope is to extend network inference methods beyond gene expression data to include other genomic and epigenomic data types.

LICENSE

The software used in constructing Predictive Networks is open source software and provided under the Apache License 2.0. All content created by the Predictive Networks application including text mining results, is provided without restriction. Data used within the PN application but derived from third party sources (such as

Pathway Commons data) are covered by their own licenses and restrictions.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figure 1 and Supplementary References [3,28,29,46–49].

ACKNOWLEDGEMENTS

We thank the Entagen software implementation team, especially Erik Bakke and James Hardwick, and Gerald Papenhausen for their efficient development of the Predictive Networks web application.

FUNDING

Funding for open access charge: National Library of Medicine of the US National Institutes of Health (grant 1R01LM010129).

Conflict of interest statement. None declared.

REFERENCES

1. Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell.*, **9**, 3273–3297.
2. Friedman,N. (2004) Inferring cellular networks using probabilistic graphical models. *Science*, **303**, 799–805.
3. Meyer,P.E., Kontos,K., Lafitte,F. and Bontempi,G. (2007) Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J. Bioinform. Syst. Biol.*, **2007**, doi:10.1155/2007/79879.
4. Margolin,A.A., Nemenman,I., Basso,K., Wiggins,C., Stolovitzky,G., Favera,R. and Califano,A. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7**, S7.
5. Faith,J.J., Hayete,B., Thaden,J.T., Mogno,I., Wierzbowski,J., Cottarel,G., Kasif,S., Collins,J.J. and Gardner,T.S. (2007) Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, **5**, e8.
6. Markowetz,F. (2010) How to understand the cell by breaking it: network analysis of gene perturbation screens. *PLoS Comput. Biol.*, **6**, e1000655.
7. Friedman,N., Linial,M., Nachman,I. and Pe'er,D. (2000) Using Bayesian networks to analyze expression data. In *RECOMB '00: Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*. ACM, New York, NY, pp. 127–135.
8. Meyer,P.E., Kontos,K. and Bontempi,G. (2007) Biological network inference using redundancy analysis. *Bioinformatics Research and Development*, Vol. 4414, Springer, Berlin/Heidelberg, pp. 16–27.
9. Djebbari,A. and Quackenbush,J. (2008) Seeded Bayesian networks: constructing genetic networks from microarray data. *BMC Syst. Biol.*, **2**, 57.
10. Aburatani,S., Goto,K., Saito,S., Fumoto,M., Imaizumi,A., Sugaya,N., Murakami,H., Sato,M., Toh,H. and Horimoto,K. (2004) ASIAN: a website for network inference. *Bioinformatics*, **20**, 2853–2856.
11. Taylor,R.C., Shah,A., Treatman,C. and Blevins,M. (2006) SEBINI: software environment for biological network inference. *Bioinformatics*, **22**, 2706–2708.
12. Haverty,P.M., Frith,M.C. and Weng,Z. (2004) CARRIE web service: automated transcriptional regulatory network inference and interactive analysis. *Nucleic Acids Res.*, **32**, W213–W216.
13. Yngvadottir,B., Macarthur,D.G., Jin,H. and Tyler-Smith,C. (2009) The promise and reality of personal genomics. *Genome Biol.*, **10**, 237.
14. Fernald,G.H., Capriotti,E., Daneshjou,R., Karczewski,K.J. and Altman,R.B. (2011) Bioinformatics challenges for personalized medicine. *Bioinformatics*, **27**, 1741–1748.
15. Mostafavi,S., Ray,D., Warde-Farley,D., Grouios,C. and Morris,Q. (2008) GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol.*, **9(Suppl. 1)**, S4.
16. Hoffmann,R. and Valencia,A. (2005) Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, **21**, ii252–ii258.
17. Ikin,A., Riveros,C., Moscato,P. and Mendes,A. (2010) The gene interaction miner: a new tool for data mining contextual information for protein-protein interaction analysis. *Bioinformatics*, **26**, 283–284.
18. GeneGo, Inc. *USA GeneGO MetaCore*. http://www.genego.com/metacore.php (31 August 2011, date last accessed).
19. Ingenuity System, Inc. *USA Ingenuity Pathway Analysis*. http://www.ingenuity.com/ (31 August 2011, date last accessed).
20. Sintek,M. and Decker,S. (2002) *TRIPLE–a query, inference, and transformation language for the semantic web*. In *International Semantic Web Conference (ISWC)*, http://triple.semanticweb.org/doc/iswc2002/abstract/ (10 November 2011, date last accessed).
21. Cerami,E.G., Gross,B.E., Demir,E., Rodchenkov,I., Babur,O., Anwar,N., Schultz,N., Bader,G.D. and Sander,C. (2010) Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res.*, **39**, D685–D690.
22. Wu,G., Feng,X. and Stein,L. (2010) A human functional protein interaction network and its application to cancer data analysis. *Genome Biol.*, **11**, R53.
23. Allias-i LingPipe. http://alias-i.com/lingpipe (31 August 2011, date last accessed).
24. Lora,V. and Gault,M.S.K.J.D. (2002) Variations in Medical Subject Headings (MeSH) mapping: from the natural language of patron terms to the controlled vocabulary of mapped lists. *J. Med. Libr. Assoc.*, **90**, 173.
25. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwoght,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
26. Himsolt,M. *Graph Modelling Language*. http://www.fim.uni-passau.de/en/fim/faculty/chairs/theoretische-informatik/projects.html (31 August 2011, date last accessed).
27. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
28. Smoot,M.E., Ono,K., Ruscheinski,J., Wang,P.-L. and Ideker,T. (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, **27**, 431–432.
29. Brown,K.R., Otasek,D., Ali,M., McGuffin,M.J., Xie,W., Devani,B., Toch,I.L.V. and Jurisica,I. (2009) NAViGaTOR: network analysis, visualization and graphing Toronto. *Bioinformatics*, **25**, 3327–3329.
30. Chickering,D.M. (1996) Learning from data: artificial intelligence and statistics V. In: Fisher,D. and Lenz,H.-J. (eds), *Learning from Data: Artificial Intelligence and Statistics V*. Springer, NY pp. 121–130.
31. Hayete,B., Gardner,T.S. and Collins,J.J. (2007) Size matters: network inference tackles the genome scale. *Mol. Syst. Biol.*, **3**, 77.
32. Haibe-Kains,B., Olsen,C., Bontempi,G. and Quackenbush,J. (2011) *Predictionet: Inference for Predictive Networks Designed for (but not limited to) Genomic Data*, https://github.com/bhaibeka/predictionet.
33. Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density

oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
34. McCall,M.N., Bolstad,B.M. and Irizarry,R.A. (2010) Frozen robust multiarray analysis (fRMA). *Biostatistics*, **11**, 242–253.
35. Steel,R.G.D. and Torrie,J.H. (1960) *Principles and procedures of statistics: with special reference to the biological sciences*. McGraw-Hill, University of Michigan, p. 481.
36. Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.
37. SpringSource. (2011) *Groovy*. http://groovy.codehaus.org/ (31 August 2011, date last accessed).
38. SpringSource. (2009) *Grails*. http://grails.org/ (31 August 2011, date last accessed).
39. Krzywinski,M., Kasaian,K., Morozova,O., Birol,I., Jones,S. and Marra,M. (2010) *Linear layout for visualization of networks. Genome Inform*, http://mkweb.bcgsc.ca/linnet (30 August 2011, date last accessed).
40. Howe,D., Costanzo,M., Fey,P., Gojobori,T., Hannick,L., Hide,W., Hill,D.P., Kania,R., Schaeffer,M., St Pierre,S. *et al.* (2008) Big data: the future of biocuration. *Nature*, **455**, 47–50.
41. Bateman,A. (2010) Curators of the world unite: the International Society of Biocuration. *Bioinformatics*, **26**, 991.
42. Fackenthal,J.D. and Olopade,O.I. (2007) Breast cancer risk associated with BRCA1 and BRCA2 in diverse populations. *Nat. Rev. Cancer*, **7**, 937–948.
43. Loi,S.M., Haibe-Kains,B., Majjaj,S., Lallemand,F., Durbecq,V., Larsimont,D., Gonzalez-Angulo,A.M., Pusztai,L., Symmans,W.F., Bardelli,A. *et al.* (2010) PIK3CA mutations associated with gene signature of low mTORC1 signaling and better outcomes in estrogen receptor-positive breast cancer. *Proc. Natl Acad. Sci. USA*, **107**, 10208–10213.
44. Culhane,A.C., Schwarzl,T., Sultana,R., Picard,K.C., Picard,S.C., Lu,T.H., Franklin,K.R., French,S.J., Papenhausen,G., Correll,M. *et al.* (2010) GeneSigDB–a curated database of gene expression signatures. *Nucleic Acids Res.*, **38**, D716–D725.
45. Loi,S.M., Haibe-Kains,B., Desmedt,C., Wirapati,P., Lallemand,F., Tutt,A.M., Gillet,C., Ellis,P., Ryder,K., Reid,J.F. *et al.* (2008) Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genomics*, **9**, 239.
46. Ding,C. and Peng,H. (2005) Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.*, **3**, 185–205.
47. McGill,W.J. (1954) Multivariate information transmission. *Psychometrika*, **9**, 97–116.
48. Meyer,P.E. (2008) Information-theoretic variable selection and network inference from microarray data. *Ph.D. Thesis*. Université Libre de Bruxelles.
49. Neapolitan,R.E. (2003) *Learning Bayesian networks*. Prentice Hall, Upper Saddle River, NJ.