Published online 2 November 2010 *Nucleic Acids Research, 2011, Vol. 39, Database issue* ***D561–D568***
doi:10.1093/nar/gkq973

The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored

Damian Szklarczyk[1], Andrea Franceschini[2], Michael Kuhn[3], Milan Simonovic[2], Alexander Roth[2], Pablo Minguez[4], Tobias Doerks[4], Manuel Stark[2], Jean Muller[5,6], Peer Bork[4,7,*], Lars J. Jensen[1,*] and Christian von Mering[2,*]

[1]Faculty of Health Sciences, Novo Nordisk Foundation Centre for Protein Research, University of Copenhagen, Denmark, [2]Faculty of Science, Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Switzerland, [3]Biotechnology Center, Technical University Dresden, [4]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany, [5]Institute of Genetics and Molecular and Cellular Biology, CNRS, INSERM, University of Strasbourg, [6]Genetic Diagnostics Laboratory, CHU Strasbourg Nouvel Hôpital Civil, Strasbourg, France and [7]Max-Delbrück-Centre for Molecular Medicine, Berlin, Germany

Received September 7, 2010; Accepted October 3, 2010

ABSTRACT

An essential prerequisite for any systems-level understanding of cellular functions is to correctly uncover and annotate all functional interactions among proteins in the cell. Toward this goal, remarkable progress has been made in recent years, both in terms of experimental measurements and computational prediction techniques. However, public efforts to collect and present protein interaction information have struggled to keep up with the pace of interaction discovery, partly because protein–protein interaction information can be error-prone and require considerable effort to annotate. Here, we present an update on the online database resource Search Tool for the Retrieval of Interacting Genes (STRING); it provides uniquely comprehensive coverage and ease of access to both experimental as well as predicted interaction information. Interactions in STRING are provided with a confidence score, and accessory information such as protein domains and 3D structures is made available, all within a stable and consistent identifier space. New features in STRING include an interactive network viewer that can cluster networks on demand, updated on-screen previews of structural information including homology models, extensive data updates and strongly improved connectivity and integration with third-party resources. Version 9.0 of STRING covers more than 1100 completely sequenced organisms; the resource can be reached at http://string-db.org.

INTRODUCTION

Proteins can form a variety of functional connections with each other, including stable complexes, metabolic pathways and a bewildering array of direct and indirect regulatory interactions. These connections can be conceptualized as networks and the size and complex organization of these networks present a unique opportunity to view a given genome as something more than just a static collection of distinct genetic functions. Indeed, the 'network view' on a genome is increasingly being taken in many areas of applied biology: protein networks are used to increase the statistical power in human genetics (1,2), to aid in drug discovery (3,4), to close gaps in metabolic enzyme knowledge (5,6) and to predict phenotypes and gene functions (7,8), to name just a few examples.

While clearly very useful, the annotation and storage of protein–protein associations in databases is less straightforward than for other types of data (such as genomic

*To whom correspondence should be addressed. Tel: +49 6221 387 8526; Fax: +49 6221 387 8517; Email: bork@embl.de
Correspondence may also be addressed to Lars J. Jensen. Tel: +45 35 32 50 25; Fax: +45 35 32 50 01; Email: lars.juhl.jensen@cpr.ku.dk
Correspondence may also be addressed to Christian von Mering. Tel: +41 44 6353147; Fax: +41 44 6356864; Email: mering@imls.uzh.ch

The authors wish it to be known that, in their opinion, the first three authors should also be regarded as joint First Authors.

sequence data or taxonomy information). This is because functional interactions between proteins can span a wide spectrum of mechanisms and specificities, often have high error rates and may depend on biological context (such as environmental condition or tissue type). Consequently, considerable information is needed to describe the various aspects of a given protein–protein association and a number of standards have been developed for this purpose with distinct levels of expressivity and specialization (9–13). Likewise, the actual annotations and interaction records themselves are scattered over a number of public resources. Experimental data on physical protein–protein interactions are mostly stored in a group of dedicated databases that together form the International Molecular Exchange (IMEx) consortium (14–21). Annotated pathway knowledge is mostly kept in a separate set of resources (22–24) and yet other interactions can be found in various organism-specific databases (25,26) or text-mining resources (27,28). Furthermore, a number of algorithms have been devised that allow *de novo* prediction of functional links between proteins (29–32), albeit usually with considerable rates of false positives and without providing hints on the specificity and type of a predicted interaction.

Given all these distinct types and sources of protein–protein association information, it is highly desirable for users to have an integration and re-appraisal that can be easily searched and browsed, at one single site. The Search Tool for the Retrieval of Interacting Genes (STRING) database resource aims to provide this service, by acting as a 'one-stop shop' for all information on functional links between proteins. It is by no means the only such site: related resources that are currently being actively maintained include VisANT (33), GeneMANIA (34), N-Browse (35), I^2D (36), APID (37), bioPIXIE (38) and ConsensusPathDB (39). Each of these sites has unique features and distinct strengths and users should carefully compare them for any specific task at hand. The main strengths of STRING lie in its unique comprehensiveness, its confidence scoring and its interactive and intuitive user interface. STRING is the only site to cover hundreds (and soon more than 1100) organisms—ranging from Bacteria and Archaea to humans. This large number of organisms, represented by their fully sequenced genomes, also enables STRING to periodically execute interaction prediction algorithms that depend on exhaustive genome sequence information. The resource also transfers interaction information between organisms where applicable, thereby significantly increasing coverage particularly for poorly studied organisms. The confidence scoring is another key feature of STRING, giving guidance to users who want to balance different levels of coverage and accuracy. Lastly, the unique and compact user interface enables fast and *ad hoc* use of the resource, with a quick learning curve and no need for setup or installation.

Here, we briefly describe the content and procedures currently used in STRING and describe new features that have been added since our last update on the resource (40).

User experience and content

Users enter STRING via its web portal (http://string-db.org) and identify one or more proteins of interest. Various types of identifiers are recognized by the system and a full-text search on gene annotations is conducted in parallel to aid in the identification. Using the search results, STRING will either recognize automatically or ask the user to disambiguate, the organism of interest. The user is then presented with the input protein(s) in the context of a graphical network of interaction partners (Figure 1). From this network, pop-up windows lead to detailed information on each node (or edge) in the network, providing accessory information on a protein or on the evidence behind a proposed connection. The network display can be modified by adding or removing proteins, changing the required confidence level and by selecting or de-selecting certain evidence types (for example, users might choose to filter out the results of computational predictions).

The interactive network viewer in STRING has been re-designed extensively. It is now based on Adobe's Flash Player (version 10 or better is recommended) and allows users to freely reposition nodes in the network. Optionally, this can be done while running a spring-embedded layout algorithm in real time. Upon switching to the 'advanced' mode of the viewer, users can also apply clustering algorithms to the network (41–43), which is then visually partitioned accordingly, in real time. All of this can be done in the context of a user-supplied background illustration; publication-ready, high-resolution image files can then be exported. Search results can also be saved in a number of abstract file formats for later use elsewhere, including the proteomics standards initiative-molecular interaction format (PSI-MI) molecular interaction standard (9). The protein information pop-up window (Figure 1, bottom) has also been re-designed using the Flash framework and now shows all available 3D structure information for a protein in the context of its domain architecture, which can be browsed interactively along the protein from N- to C-terminus. Apart from PDB entries, the structure information now also includes pre-computed homology models, made available via a collaboration with the SwissModel repository (44).

The current extent of protein–protein association information in STRING is summarized in Figure 2. The majority of associations actually derive from predictions—either from prediction algorithms that are based on analyzing genomic information ('genomic context'-methods) or from transferring associations/interactions between organisms ('interolog'-transfer). Importantly, all associations in STRING are provided with a probabilistic confidence score, which is derived by separately benchmarking groups of associations against the manually curated functional classification scheme of the KEGG database (22). Each score represents a rough estimate of how likely a given association describes a functional linkage between two proteins that is at least as specific as that between an average pair of proteins annotated on the same 'map' or 'pathway'

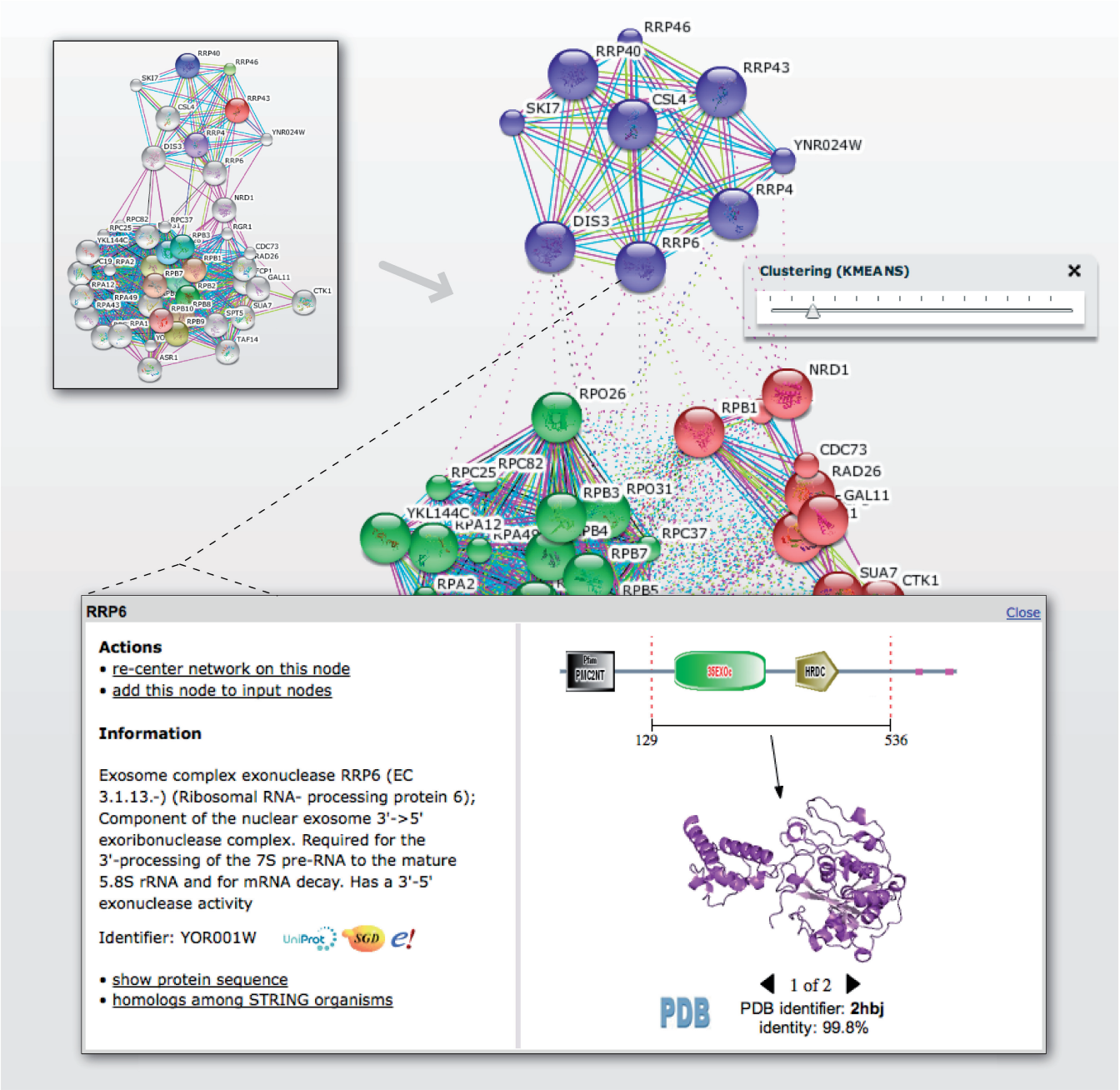

Figure 1. Protein network visualization on the STRING website. The figure shows a composite of two screenshots, illustrating a typical user interaction with STRING (focused on a specific protein network in *Saccharomyces cerevisiae*). Upon querying the database with four yeast proteins, the resource first reports a raw network consisting of the highest scoring interaction partners (upper left corner). This network can then be rearranged and clustered directly in the browser window revealing tightly connected functional modules (arrow). For each interaction (or protein), additional information is accessible via dedicated pop-up windows; the bottom part of the figure shows an exemplary pop-up with the information regarding a specific yeast protein.

in KEGG. The various major sources of interaction/association data in STRING are benchmarked independently; a combined score is computed which indicates higher confidence when more than one type of information supports a given association. All scores and association data in STRING are pre-computed and are also available for wholesale download (free for non-profit institutions). Fully sequenced genomes in STRING are imported from RefSeq (45) and Ensembl (46), as well as from a number of dedicated sites, and are hand-screened for completeness and non-redundancy. For this large space of complete genomes, STRING also stores the results of exhaustive cross-genome homology searches, in order to be able to transfer interactions among organisms. As of version 9.0, this extensive body of protein–protein similarity data is imported from and cross-linked with the Similarity Matrix of Proteins (SIMAP) project (47).

	E. coli	*S. cerevisiae*	*H. sapiens*	STRING total
predicted associations				
via gene neighborhood	14,685	0	0	6.9 Mio
via gene fusions	2,930	1,127	4,017	1.3 Mio
via gene co-occurrence	148,420	1,046	7,583	33.0 Mio
via gene co-expression	56,397	58,848	26,278	674,416
known associations				
textmining (co-occurrence)	26,796	70,760	5.5 Mio	7.3 Mio
textmining (natural language processing)	1,000	3,155	226,336	274,214
BIND	368	3,014	2,284	16,741
BioCarta	0	0	12,761	12,761
BioCyc	2,843	809	969	160,462
BioGrid	0	297,276	30,443	365,156
DIP	1,433	12,429	1,088	23,589
Gene Ontology protein complexes	82	6,980	8,052	35,190
HPRD	0	0	18,848	18,848
INTACT	182,33	75,042	28,562	182,524
KEGG	4,965	6,746	38,810	1.9 Mio
MINT	117	66,083	18,839	126,659
NCI Nature / Pathway Interaction DB	0	0	18,909	18,909
PDB	18,586	5,310	2,443	243,434
Reactome	0	83,855	262,482	2.8 Mio
associations transferred from other organisms				
from predicted associations	142,778	39,744	45,890	44.1 Mio
from known associations	21,029	20,152	56,386	12.8 Mio

Figure 2. Association counts and data sources. The table shows the number of pair-wise protein–protein associations processed for STRING (version 8.3), listed separately for three important model organisms as well as for the database as a whole. The associations are counted non-directionally, i.e. protein pairs A–B and B–A are counted only once. Identical associations reported by different sources are counted separately under each source, unless they can be traced to the very same publication record and have been imported from primary interaction databases (in case several such databases agree on an interaction, it is arbitrarily counted for only one of them).

It should be stressed that interactions in STRING are not limited to direct, physical interactions between two proteins. Instead, proteins may also be linked because, for example, they exhibit a genetic interaction or are known to catalyze subsequent steps in a metabolic pathway. Most associations, especially when derived from one of the prediction algorithms, currently can neither be specified with much precision in terms of their mode of interaction, nor in terms of the cellular conditions under which they occur (e.g. development time points, environmental conditions, specific cell types, etc.). Because of this, the fundamental unit stored in STRING is the 'functional association', i.e. the specific and biologically meaningful functional connection between two proteins. Within this definition, STRING aims to uncover the entire space of 'possible' interactions for any fully sequenced organism; it is likely that only a subset of these interactions will be realized in any given cell. The number of interactions stored in STRING has grown considerably over the years and is projected to grow further as more information becomes available. Previous versions of the resource are kept accessible online, such that studies that refer to a given version of STRING can later be reproduced.

Integration with other resources

One central aim of the STRING project is to achieve and maintain cross-connectivity and integration with other public resources in a user-friendly manner. Apart from making the entire SQL database back-end available for download (free for non-profit institutions), this is mainly achieved via the following routes:

First, the database maintains mutual HTML cross-references with a number of widely used websites, including UniProt (48), SMART (49), GeneCards (50) and SwissModelRepository (44). Notably, such cross references do not have to be limited to simple text-based HTML links. Instead, partner websites can embed minimized icon-previews of STRING networks within their own web pages, using the capabilities of STRINGs API interface (as described in the last update) (40). The SMART and SwissModelRepository sites already

use this option, requesting the network preview images—when needed, at run time—based on pre-determined name-space mappings. Such embedded previews do not have to be limited to static images; external sites can also provide pop-up windows for any protein of interest, the content of the pop-up is then provided by STRING [variants of this mechanism are currently used by the resources Reflect (51) and ViralZone (http://expasy.org/viralzone)]. As another new feature of the user interface, permanent URLs can now be retrieved for almost all pages served by STRING—this facilitates cross-linking and archiving and also indexing by search engines and meta-sites.

Second, partner websites can choose to embed the entire STRING website into their own pages (52,53), for example, using HTML inline frames (iframes). A notable example for this is the BioGPS Community Gene Portal System (53); this site provides 'plugins' through which users can connect any number of external websites into freely configurable screen layouts. A STRING plugin has been established at BioGPS; it is currently among the most frequently used plugins there.

Third, users can choose to work with STRING networks from inside the Cytoscape software. Cytoscape is a widely used open-source software framework for network visualization and manipulation (54,55); it can be very flexibly extended, with a rapidly growing number of network-centered manipulation and analysis tools. There are several options for loading STRING data into Cytoscape: users can save a given network from the STRING site to a local file, which can then be opened by Cytoscape (preferably using the PSI-MI format). Users can also query STRING directly from within Cytoscape; this is made possible via a dedicated plugin 'StringWSClient' that exposes much of the STRING query interface, including organism disambiguation. Lastly, the perhaps most important way to query STRING from within Cytoscape is via the 'PSICQUIC' query interface ('PSICQUIC Web Service Universal Client' in Cytoscape). PSICQUIC is a newly developed

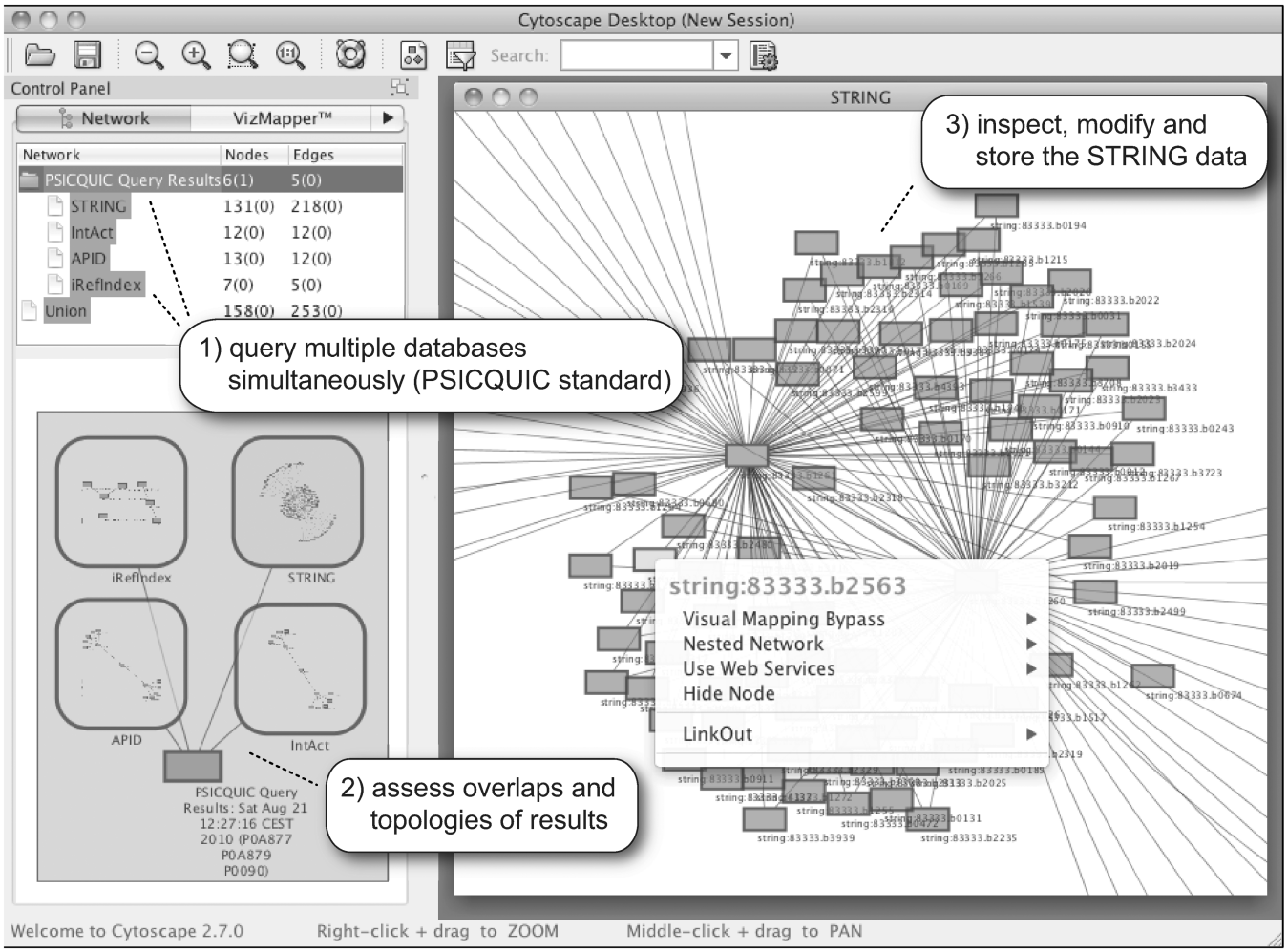

Figure 3. Accessing STRING data from within Cytoscape. Two proteins from *Escherichia coli* were used as queries for the 'PSICQUIC Web Service Universal Client' import-plugin of Cytoscape. Multiple databases have reported hits for these queries (upper left panel); in this case STRING has reported the largest number of hits. The resulting four networks are largely non-overlapping, both in terms of name-spaces as well as in terms of the actual interactors reported. The imported STRING network (right) is shown in detail; it can be used as the basis of further refinement, post-processing and analysis in Cytoscape.

standard that allows interaction queries across a growing number of compliant database resources (56); STRING has implemented this standard as of version 8.3 and can thus now be queried directly alongside a number of other resources (Figure 3).

Lastly, a new call-back interface allows STRING to be 'branded' by third-party resources, who may wish to project their own information onto the STRING name space and thereby onto the STRING network data (Figure 4). This allows such resources to take advantage of the extensive user-interface features of STRING, as well as tapping into the existing user base, with very little additional coding effort of their own. This mechanism requires no specific setup on the STRING side—instead, our resource is simply instructed to query the third-party site at runtime, for any additional information that is to be displayed alongside the STRING network. Data updates at the STRING site are usually accommodated automatically, since the name space itself is changed only at the major release updates.

Published use cases

STRING has been used in projects of various scales—both in large, organism-wide studies but also in focused projects that are restricted to a few proteins or to a single pathway only. Studies of the latter type often make use of STRING as a discovery tool, taking advantage of the pre-computed and confidence-scored association predictions that it provides. Examples include the discoveries of a missing enzyme in Bacillothiol biosynthesis in Bacilli (57), of a previously unknown chaperone subunit in Cytochrome C oxidase assembly (58) or of a missing enzyme in uric acid degradation in mammals (59).

Another way to use STRING is to download and extend its relational database schema; this can, for example, be useful for projects dedicated to additional types of information (e.g. small molecule interactors in the case of our partner project STITCH) (60) or for projects wishing to rely on a single source of completely sequenced genomes with associated homology data (e.g. in the case of the gene orthology resource eggNOG) (61). Users not wishing to download and install the entire database schema have the alternative to download compact flat-files; these contain only the actual interaction information or information regarding the interacting proteins themselves (sequences, identifiers, etc.).

A unique strength of STRING lies in its comprehensiveness, albeit at the expense of considerable false-positive rates. Because of this, organism-wide studies represent perhaps the most interesting use cases and they are probably best done when they involve integration of orthogonal data types (since this may allow the noise in both data sets to cancel out). Examples include the filtering and extension of results from large-scale genetic screens (62,63) or the annotation of large groups of proteins having a specific post-translational modification (64). Another intriguing application scenario is to use STRING for search-space reduction in epistasis screens. This is done under the assumption that gene loci showing genetic epistasis should also often show up as functionally linked in STRING. Indeed, this approach has been demonstrated to work on human association mapping data, providing the statistical power to link up loci that show a non-additive effect when mutated together (1,2). Approaches such as this are expected to gain further power, as the information in STRING becomes even more comprehensive and precise in future updates.

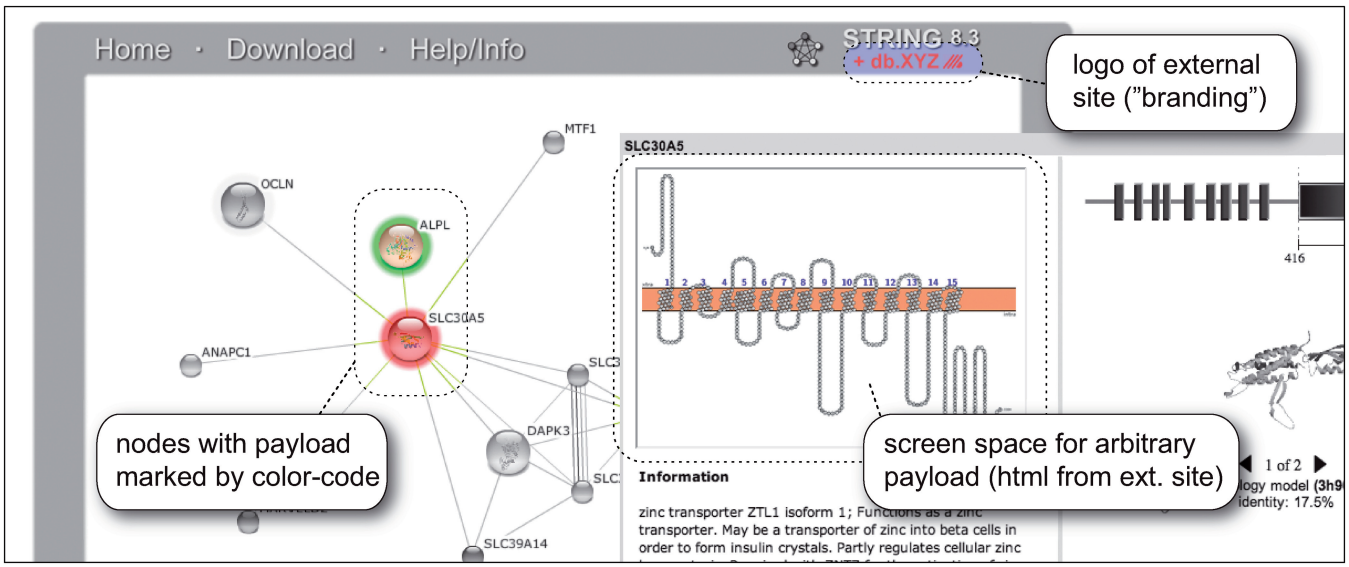

Figure 4. Projecting third-party data onto the STRING web-surface. STRING provides a consistent name space that encompasses genes, genomes, protein and interaction networks, all of which can be easily searched and browsed. These features can now be employed by external web-resources, via a simple call-back mechanism. External resources can provide cross-links to STRING, together with a call-back address capable of serving a simple text-based interface protocol. At run-time, STRING will then automatically call the external site and project arbitrary 'payload' information onto the protein network that is being browsed. The figure shows a fictitious example scenario, served from an in-house test server. As of version 9.0, STRING will also be able to accept protein–protein connections as payload, showing them in a dedicated 'evidence channel' distinct from the seven built-in channels. Implementation details are available in the online documentation.

ACKNOWLEDGEMENTS

The authors wish to thank the PSICQUIC consortium for early access to their standardization effort, and Dr Gary Bader for technical help with the Cytoscape plugin.

FUNDING

STRING is funded by the Swiss Institute of Bioinformatics, by the Novo Nordisk Foundation Center for Protein Research and by the European Molecular Biology Laboratory (EMBL). Funding for open access charges: University of Zurich, through its Research Priority program 'Systems Biology and Functional Genomics'.

Conflict of interest statement. None declared.

REFERENCES

1. Pattin,K.A. and Moore,J.H. (2009) Role for protein-protein interaction databases in human genetics. *Expert Rev. Proteomics*, **6**, 647–659.
2. Emily,M., Mailund,T., Hein,J., Schauser,L. and Schierup,M.H. (2009) Using biological networks to search for interacting loci in genome-wide association studies. *Eur. J. Hum. Genet.*, **17**, 1231–1240.
3. Pujol,A., Mosca,R., Farres,J. and Aloy,P. (2010) Unveiling the role of network and systems biology in drug discovery. *Trends Pharmacol. Sci.*, **31**, 115–123.
4. Klipp,E., Wade,R.C. and Kummer,U. (2010) Biochemical network-based drug-target prediction. *Curr. Opin. Biotechnol.*, **21**, 511–516.
5. Janga,S.C., Diaz-Mejia,J.J. and Moreno-Hagelsieb,G. (2010) Network-based function prediction and interactomics: The case for metabolic enzymes. *Metab. Eng.*
6. Orth,J.D. and Palsson,B.O. (2010) Systematizing the generation of missing metabolic knowledge. *Biotechnol. Bioeng.*, **107**, 403–412.
7. Wang,P.I. and Marcotte,E.M. (2010) It's the machine that matters: Predicting gene function and phenotype from protein networks. *J. Proteomics*, **73**, 2277–2289.
8. Lage,K., Karlberg,E.O., Storling,Z.M., Olason,P.I., Pedersen,A.G., Rigina,O., Hinsby,A.M., Tumer,Z., Pociot,F., Tommerup,N. *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.*, **25**, 309–316.
9. Hermjakob,H., Montecchi-Palazzi,L., Bader,G., Wojcik,J., Salwinski,L., Ceol,A., Moore,S., Orchard,S., Sarkans,U., von Mering,C. *et al.* (2004) The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nat. Biotechnol.*, **22**, 177–183.
10. Luciano,J.S. (2005) PAX of mind for pathway researchers. *Drug Discov. Today*, **10**, 937–942.
11. Hucka,M., Finney,A., Sauro,H.M., Bolouri,H., Doyle,J.C., Kitano,H., Arkin,A.P., Bornstein,B.J., Bray,D., Cornish-Bowden,A. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–531.
12. Le Novere,N., Hucka,M., Mi,H., Moodie,S., Schreiber,F., Sorokin,A., Demir,E., Wegner,K., Aladjem,M.I., Wimalaratne,S.M. *et al.* (2009) The Systems Biology Graphical Notation. *Nat. Biotechnol.*, **27**, 735–741.
13. Lloyd,C.M., Halstead,M.D. and Nielsen,P.F. (2004) CellML: its future, present and past. *Prog. Biophys. Mol. Biol.*, **85**, 433–450.
14. Orchard,S., Kerrien,S., Jones,P., Ceol,A., Chatr-Aryamontri,A., Salwinski,L., Nerothin,J. and Hermjakob,H. (2007) Submit your interaction data the IMEx way: a step by step guide to trouble-free deposition. *Proteomics*, **7(Suppl 1)**, 28–34.
15. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
16. Aranda,B., Achuthan,P., Alam-Faruque,Y., Armean,I., Bridge,A., Derow,C., Feuermann,M., Ghanbarian,A.T., Kerrien,S., Khadake,J. *et al.* (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38**, D525–D531.
17. Ceol,A., Chatr Aryamontri,A., Licata,L., Peluso,D., Briganti,L., Perfetto,L., Castagnoli,L. and Cesareni,G. (2010) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.*, **38**, D532–D539.
18. Guldener,U., Munsterkotter,M., Oesterheld,M., Pagel,P., Ruepp,A., Mewes,H.W. and Stumpflen,V. (2006) MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res.*, **34**, D436–D441.
19. Chautard,E., Ballut,L., Thierry-Mieg,N. and Ricard-Blum,S. (2009) MatrixDB, a database focused on extracellular protein-protein and protein-carbohydrate interactions. *Bioinformatics*, **25**, 690–691.
20. Goll,J., Rajagopala,S.V., Shiau,S.C., Wu,H., Lamb,B.T. and Uetz,P. (2008) MPIDB: the microbial protein interaction database. *Bioinformatics*, **24**, 1743–1744.
21. Breitkreutz,B.J., Stark,C., Reguly,T., Boucher,L., Breitkreutz,A., Livstone,M., Oughtred,R., Lackner,D.H., Bahler,J., Wood,V. *et al.* (2008) The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res.*, **36**, D637–D640.
22. Kanehisa,M., Goto,S., Furumichi,M., Tanabe,M. and Hirakawa,M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.
23. Matthews,L., Gopinath,G., Gillespie,M., Caudy,M., Croft,D., de Bono,B., Garapati,P., Hemish,J., Hermjakob,H., Jassal,B. *et al.* (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, **37**, D619–D622.
24. Karp,P.D., Ouzounis,C.A., Moore-Kochlacs,C., Goldovsky,L., Kaipa,P., Ahren,D., Tsoka,S., Darzentas,N., Kunin,V. and Lopez-Bigas,N. (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.*, **33**, 6083–6089.
25. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
26. Tweedie,S., Ashburner,M., Falls,K., Leyland,P., McQuilton,P., Marygold,S., Millburn,G., Osumi-Sutherland,D., Schroeder,A., Seal,R. *et al.* (2009) FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res.*, **37**, D555–D559.
27. Rodriguez-Esteban,R. (2009) Biomedical text mining and its applications. *PLoS Comput. Biol.*, **5**, e1000597.
28. Hoffmann,R. and Valencia,A. (2004) A gene network for navigating the literature. *Nat. Genet.*, **36**, 664.
29. Lewis,A.C., Saeed,R. and Deane,C.M. (2010) Predicting protein-protein interactions in the context of protein evolution. *Mol. Biosyst.*, **6**, 55–64.
30. Skrabanek,L., Saini,H.K., Bader,G.D. and Enright,A.J. (2008) Computational prediction of protein-protein interactions. *Mol. Biotechnol.*, **38**, 1–17.
31. Huynen,M.A., Snel,B., von Mering,C. and Bork,P. (2003) Function prediction and protein networks. *Curr. Opin. Cell Biol.*, **15**, 191–198.
32. Valencia,A. and Pazos,F. (2002) Computational methods for the prediction of protein interactions. *Curr. Opin. Struct. Biol.*, **12**, 368–373.
33. Hu,Z., Hung,J.H., Wang,Y., Chang,Y.C., Huang,C.L., Huyck,M. and DeLisi,C. (2009) VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. *Nucleic Acids Res.*, **37**, W115–W121.
34. Warde-Farley,D., Donaldson,S.L., Comes,O., Zuberi,K., Badrawi,R., Chao,P., Franz,M., Grouios,C., Kazi,F., Lopes,C.T. *et al.* (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.*, **38(Suppl)**, W214–W220.
35. Kao,H.L. and Gunsalus,K.C. (2008) Browsing multidimensional molecular networks with the generic network browser (N-Browse). *Curr. Protoc. Bioinformatics*, **Chapter 9**, Unit 9 11.

36. Niu,Y., Otasek,D. and Jurisica,I. (2010) Evaluation of linguistic features useful in extraction of interactions from PubMed; application to annotating known, high-throughput and predicted interactions in I2D. *Bioinformatics*, **26**, 111–119.
37. Prieto,C. and De Las Rivas,J. (2006) APID: Agile Protein Interaction DataAnalyzer. *Nucleic Acids Res.*, **34**, W298–W302.
38. Myers,C.L., Chiriac,C. and Troyanskaya,O.G. (2009) Discovering biological networks from diverse functional genomic data. *Methods Mol. Biol.*, **563**, 157–175.
39. Kamburov,A., Wierling,C., Lehrach,H. and Herwig,R. (2009) ConsensusPathDB–a database for integrating human functional interaction networks. *Nucleic Acids Res.*, **37**, D623–D628.
40. Jensen,L.J., Kuhn,M., Stark,M., Chaffron,S., Creevey,C., Muller,J., Doerks,T., Julien,P., Roth,A., Simonovic,M. *et al.* (2009) STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412–D416.
41. Pavlopoulos,G.A., Moschopoulos,C.N., Hooper,S.D., Schneider,R. and Kossida,S. (2009) jClust: a clustering and visualization toolbox. *Bioinformatics*, **25**, 1994–1996.
42. de Hoon,M.J., Imoto,S., Nolan,J. and Miyano,S. (2004) Open source clustering software. *Bioinformatics*, **20**, 1453–1454.
43. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
44. Kiefer,F., Arnold,K., Kunzli,M., Bordoli,L. and Schwede,T. (2009) The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res.*, **37**, D387–D392.
45. Pruitt,K.D., Tatusova,T., Klimke,W. and Maglott,D.R. (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.*, **37**, D32–D36.
46. Hubbard,T.J., Aken,B.L., Ayling,S., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Clarke,L. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.
47. Rattei,T., Tischler,P., Gotz,S., Jehl,M.A., Hoser,J., Arnold,R., Conesa,A. and Mewes,H.W. (2010) SIMAP–a comprehensive database of pre-calculated protein sequence similarities, domains, annotations and clusters. *Nucleic Acids Res.*, **38**, D223–D226.
48. Apweiler,R., Martin,M.J., O'Donovan,C., Magrane,M., Alam-Faruque,Y., Antunes,R., Barell,D., Bely,B., Bingley,M., Binns,D. *et al.* (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.
49. Letunic,I., Doerks,T. and Bork,P. (2009) SMART 6: recent updates and new developments. *Nucleic Acids Res.*, **37**, D229–D232.
50. Safran,M., Dalah,I., Alexander,J., Rosen,N., Iny Stein,T., Shmoish,M., Nativ,N., Bahir,I., Doniger,T., Krug,H. *et al.* (2010) GeneCards Version 3: the human gene integrator. *Database*, **2010**, baq020.
51. Pafilis,E., O'Donoghue,S.I., Jensen,L.J., Horn,H., Kuhn,M., Brown,N.P. and Schneider,R. (2009) Reflect: augmented browsing for the life scientist. *Nat. Biotechnol.*, **27**, 508–510.
52. Liebel,U., Kindler,B. and Pepperkok,R. (2005) Bioinformatic "Harvester": a search engine for genome-wide human, mouse, and rat protein resources. *Methods Enzymol.*, **404**, 19–26.
53. Wu,C., Orozco,C., Boyer,J., Leglise,M., Goodale,J., Batalov,S., Hodge,C.L., Haase,J., Janes,J., Huss,J.W. 3rd *et al.* (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.*, **10**, R130.
54. 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.
55. Cline,M.S., Smoot,M., Cerami,E., Kuchinsky,A., Landys,N., Workman,C., Christmas,R., Avila-Campilo,I., Creech,M., Gross,B. *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.*, **2**, 2366–2382.
56. Orchard,S., Albar,J.P., Deutsch,E.W., Eisenacher,M., Binz,P.A. and Hermjakob,H. (2010) implementing data standards: a report on the HUPOPSI workshop September 2009, Toronto, Canada. *Proteomics*, **10**, 1895–1898.
57. Gaballa,A., Newton,G.L., Antelmann,H., Parsonage,D., Upton,H., Rawat,M., Claiborne,A., Fahey,R.C. and Helmann,J.D. (2010) Biosynthesis and functions of bacillithiol, a major low-molecular-weight thiol in Bacilli. *Proc. Natl Acad. Sci. USA*, **107**, 6482–6486.
58. Banci,L., Bertini,I., Ciofi-Baffoni,S., Katsari,E., Katsaros,N., Kubicek,K. and Mangani,S. (2005) A copper(I) protein possibly involved in the assembly of CuA center of bacterial cytochrome c oxidase. *Proc. Natl Acad. Sci. USA*, **102**, 3994–3999.
59. Ramazzina,I., Folli,C., Secchi,A., Berni,R. and Percudani,R. (2006) Completing the uric acid degradation pathway through phylogenetic comparison of whole genomes. *Nat. Chem. Biol.*, **2**, 144–148.
60. Kuhn,M., Szklarczyk,D., Franceschini,A., Campillos,M., von Mering,C., Jensen,L.J., Beyer,A. and Bork,P. (2010) STITCH 2: an interaction network database for small molecules and proteins. *Nucleic Acids Res.*, **38**, D552–D556.
61. Muller,J., Szklarczyk,D., Julien,P., Letunic,I., Roth,A., Kuhn,M., Powell,S., von Mering,C., Doerks,T., Jensen,L.J. *et al.* (2010) eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res.*, **38**, D190–D195.
62. Wang,L., Tu,Z. and Sun,F. (2009) A network-based integrative approach to prioritize reliable hits from multiple genome-wide RNAi screens in Drosophila. *BMC Genomics*, **10**, 220.
63. Mummery-Widmer,J.L., Yamazaki,M., Stoeger,T., Novatchkova,M., Bhalerao,S., Chen,D., Dietzl,G., Dickson,B.J. and Knoblich,J.A. (2009) Genome-wide analysis of Notch signalling in Drosophila by transgenic RNAi. *Nature*, **458**, 987–992.
64. Choudhary,C., Kumar,C., Gnad,F., Nielsen,M.L., Rehman,M., Walther,T.C., Olsen,J.V. and Mann,M. (2009) Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science*, **325**, 834–840.