STRING 8—a global view on proteins and their functional interactions in 630 organisms

Lars J. Jensen¹,², Michael Kuhn¹, Manuel Stark³, Samuel Chaffron³, Chris Creevey¹, Jean Muller¹, Tobias Doerks¹, Philippe Julien⁴, Alexander Roth³, Milan Simonovic³, Peer Bork¹,⁵,* and Christian von Mering³

¹European Molecular Biology Laboratory, Heidelberg, Germany, ²Novo Nordisk Foundation Centre for Protein Research, University of Copenhagen, Denmark, ³Institute of Molecular Biology and Swiss Institute of Bioinformatics, University of Zurich, ⁴Centre for Integrative Genomics, University of Lausanne, Switzerland and ⁵Max-Delbrück-Centre for Molecular Medicine, Berlin, Germany

Received September 14, 2008; Accepted October 6, 2008

ABSTRACT

Functional partnerships between proteins are at the core of complex cellular phenotypes, and the networks formed by interacting proteins provide researchers with crucial scaffolds for modeling, data reduction and annotation. STRING is a database and web resource dedicated to protein–protein interactions, including both physical and functional interactions. It weights and integrates information from numerous sources, including experimental repositories, computational prediction methods and public text collections, thus acting as a meta-database that maps all interaction evidence onto a common set of genomes and proteins. The most important new developments in STRING 8 over previous releases include a URL-based programming interface, which can be used to query STRING from other resources, improved interaction prediction via genomic neighborhood in prokaryotes, and the inclusion of protein structures. Version 8.0 of STRING covers about 2.5 million proteins from 630 organisms, providing the most comprehensive view on protein–protein interactions currently available. STRING can be reached at http://string-db.org/.

EXTENDING THE SOURCES OF INTERACTION INFORMATION

The basic interaction unit in STRING is the ‘functional association’, which is defined in this database as the specific and meaningful interaction between two proteins that jointly contribute to the same functional process. With respect to the interacting proteins, STRING does not consider any specific splicing isoforms or posttranslational modifications, but instead represents each protein-coding experimental techniques for their elucidation are diverse, often not directly comparable, and less reliable than genome sequencing. Nevertheless, protein–protein interaction networks (or also ‘association networks’ in case functional associations are included) are a crucial ingredient for any system-level understanding of cellular machineries (1–5). Furthermore, protein networks can serve very concrete, practical purposes such as filtering and assessing high-throughput functional genomics data, and providing intuitive visual scaffolds for annotating the structural, functional and evolutionary properties of proteins.

The database and web-tool STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is a metasource that aggregates most of the available information on protein–protein associations, scores and weights it, and augments it with predicted interactions, as well as with the results of automatic literature-mining searches. Since its first release in 2000 (6), it has grown into the most comprehensive resource of its type. It builds upon and extends the excellent, manual annotation efforts undertaken at primary protein interaction databases (7–12) and at databases of curated pathway knowledge (13–15). Here, we describe new features that have been added since our report on the previous release, STRING 7 (16).
locus in a genome by a single protein (the longest iso-
form). Thus, and because STRING aggregates data and
predictions stemming from a wide spectrum of cell types
and environmental conditions, it aims to represent the
union of all possible protein–protein links. From this
union, the actual network for any given spatio-temporal
snapshot of the cell can in principle be deduced by projec-
tion, for example by removing proteins known to be not
expressed or not active under the conditions studied (17).

In keeping with the above definitions, STRING imports
protein association knowledge not only from databases of
physical interactions, but also from databases of curated
biological pathway knowledge. Apart from the resources
already included in the previous release [MINT (10),
HPRD (9), BIND (12), DIP (11), BioGRID (8), KEGG
(13) and Reactome (14)], a number of resources have been
newly included [IntAct (7), EcoCyc (15), NCI-Nature
Pathway Interaction Database and Gene Ontology (GO)
protein complexes]. For the full STRING release, this
set of previously known and well-described interactions
is then complemented by interactions that are predicted
computationally, specifically for STRING, using a
number of prediction algorithms (18,19). First, we con-
duct systematic searches for genes that are found in
close proximity within prokaryotic chromosomes, which
is a good indicator for functional linkage. Second, we
search for instances where genes have joined to encode a
single fusion protein, which is indicative of functional
linkage even in organisms where the two proteins have
not fused. Third, we search for gene families that share
above-random similarities in their evolutionary histories
(i.e. they have similar ‘phylogenetic profiles’). This,
again, predicts that they contribute to similar functional
processes in the cell. Fourth, we conduct searches for
genes that display a similar transcriptional response
across a variety of conditions (co-expression). Individ-
ually, the above predictors may not always have the
specificity of direct experimental interaction assays; how-
ever, when used in concert and integrated probabilisti-
cally, the performance even of relatively weak predictors
can rival that of experimental data (20).

Lastly, two further sources of interactions in STRING
are actually providing the majority of associations; these
are text-mining and interaction transfer between organ-
isms. For the former, we parse a large body of scientific
texts [SGD (21), OMIM (22), The Interactive Fly, and
all abstracts from PubMed]. We search for statistically
relevant co-occurrences of gene names, and also extract
a subset of semantically specified interactions using
Natural Language Processing (23). For the transfer of
interactions between organisms, we estimate whether a
pair of interacting proteins found conserved in another
organism justifies the transfer of the interaction to that
other organism (24). The transferred interactions, as well
as all predicted or imported interactions, are benchmarked
and scored against a common reference of functional par-
tnership [we currently use the joint membership of proteins
in biological pathways, as annotated at KEGG (13),
as our gold-standard].

Together, the above sources of interactions, including
predictions and transfers, result in a uniquely high
coverage of the interaction networks stored in STRING
(Figure 1), particularly for well-studied model organisms.
Since the previous release, STRING has almost doubled
the number of supported organisms, which now stands at
630. The number of stored interactions has increased as
well, to a total of more than 50 million. Since the various
subtypes of the interaction evidence are stored separately
in the database, they can be disabled at will—giving users
the ability to adjust the scope and specificity of STRING
towards their particular application.

EXTENDED DEFINITION OF CONSERVED
GENOMIC NEIGHBORHOOD

When working with prokaryotes, scientists have long
used conserved genomic neighborhood arrangements
of genes to infer functional linkage, assuming that such
arrangements reflect polycistronic transcription units
(operons). STRING has followed this principle, compiling
and benchmarking protein–protein associations based
on close, co-directional neighborhood of genes on the
gene. As of version 8, this has been extended to cover also
neighboring genes that are counter-directional in a
head-to-head orientation (‘divergent transcription’). Such
divergently oriented gene pairs have been shown to be
indicative of functional linkage as well (25), albeit with
somewhat lower confidence. Often, one of the two genes
is a transcriptional regulator, targeting the neighboring
gene (25). STRING now uses this type of arrangement
in its neighborhood algorithm as well (benchmarking sepa-
 radically, Figure 2). In addition, STRING is now more error
tolerant when assembling conserved neighborhoods,
ignoring short, partially overlapping genes on the anti-
sense strand that are likely to be spurious predictions.

INTEGRATION OF PROTEIN STRUCTURES

For each update, STRING now parses all entries of the
PDB database of protein structures (26). The use of pro-
tein structures is two-fold: first, to inform the user that a
given protein—or a close homolog thereof—indeed has
3D structure information. In this case, a small preview of
a representative structure is shown in the network,
and the user can follow it to view the full structure and
to proceed to the PDB website. Second, protein structures
serve as interaction evidence themselves, when more than
one distinct peptide chain is found in the structure. In this
case, a stable and reliable protein–protein interaction is
assumed.

NEW PROGRAMMING INTERFACE

To facilitate the integration of STRING into network
tools like Cytoscape (27) and workflow engines like
Taverna (28), we have created an application program-
ing interface (API) that allows access to the interaction
network in computer-readable formats (Figure 3). Addi-
tionally, specific API functions allow retrieval of individ-
ual records from our database, for example to map a
protein via its name onto a STRING entry. We further
envision that the STRING API will be useful to developers of web services, who plan to make use of the STRING interaction network. If a particular web service needs access to the complete set of interactions, it may still be advisable to maintain a local copy of our data distribution. However, if the service requires access to many different subsets (depending on user input), querying STRING via its API could reduce administrative load.

The API is called by constructing a URL that contains the type of the request, the desired output format and the input items. The STRING server then returns the result of the computation in the desired format. Further documentation can be accessed via the STRING homepage.

**USE SCENARIOS**

Apart from the ad hoc and barrier-free access through the website, STRING can be downloaded and used locally, either in the form of concise flat-files or as a mirror...
installation of the complete relational database back-end (some of the downloads do require a free, nonredistribution license applicable to academic nonprofit users). The interacting entities in STRING can be set to be either proteins, or groups of orthologs spanning multiple organisms (‘COG-mode’). For the latter, STRING relies on an updated and extended version of the COGs ['Clusters of Orthologous Groups' (29)], which is being maintained at the eggNOG database (30). A variety of other databases use STRING networks as a basis for further computations/annotations, for example by augmenting the networks with small molecules [STITCH, (31)], or by using the network to increase the power of kinase–substrate predictions [NetworKIN, (32)].

ACKNOWLEDGEMENTS

The authors wish to thank Dianna Fisk from the Saccharomyces Genome Database, and Thomas B. Brody from The Interactive Fly, for access to gene summary paragraphs. Code development was partially conducted at the ‘WebService BioHackathon 2008’ in Tokyo, Japan.

FUNDING

Swiss Institute of Bioinformatics; University of Zurich through its Research Priority Program ‘Systems Biology and Functional Genomics’; European Commission’s FP6 Programme through the ADIT Integrated Project (LSHB-CT-2005-511065); BioSapiens Network of Excellence (LSHG-CT-2003-503265). Funding for open access charge: University of Zurich.

REFERENCES

1. Bader, S., Kuhn, S. and Gavin, A.C. (2008) Interaction networks for systems biology. FEBS Lett., 582, 1220–1224.
2. Devos, D. and Russell, R.B. (2007) A more complete, complexed and structured interactome. Curr. Opin. Struct. Biol., 17, 370–377.
3. Hu, Z., Mellor, J., Wu, J., Kanehisa, M., Stuart, J.M. and DeLisi, C. (2007) Towards zoomable multidimensional maps of the cell. Nat. Biotechnol., 25, 547–554.
4. Christensen, C., Thakar, J. and Albert, R. (2007) Systems-level insights into cellular regulation: inferring, analysing, and modelling intracellular networks. IET Syst. Biol., 1, 61–77.
5. Kohler, S., Bauer, S., Horn, D. and Robinson, P.N. (2008) Walking the interactome for prioritization of candidate disease genes. Am. J. Hum. Genet., 82, 949–958.
6. Snel, B., Lehmann, G., Bork, P. and Huynen, M.A. (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res., 28, 3442–3444.
7. Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R. et al. (2007) IntAct—open source resource for molecular interaction data. Nucleic Acids Res., 35, D561–D565.
8. 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.
9. Mishra, G.R., Suresh, M., Kumaran, K., Kannabiran, N., Suresh, S., Bala, P., Shivakumar, K., Anuradha, N., Reddy, R., Raghavan, T.M. et al. (2006) Human protein reference database—2006 update. Nucleic Acids Res., 34, D411–D414.
10. Chatr-aryamontri, A., Ceol, A., Palazzi, L.M., Nardelli, G., Schneider, M.V., Castagnoli, L. and Cesareni, G. (2007) MINT: the Molecular INTeraction database. Nucleic Acids Res., 35, D572–D574.
11. 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.
12. Alfarano, C., Andreade, C.E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutiler, K., Burgess, E. et al. (2005) The Biomolecular Interaction Network Database and related tools 2005. Nucleic Acids Res., 33, D418–D424.

13. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. et al. (2008) KEGG for linking genomes to life and the environment. Nucleic Acids Res., 36, D480–D484.

14. Von Mering, C., Jensen, L.J., Kuhn, M., Chaffron, S., Doerks, T., de Lichtenberg, U., Jensen, L.J., Brunak, S. and Bork, P. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res., 33, D433–D437.

15. Korbel, J.O., Jensen, L.J., von Mering, C. and Bork, P. (2004) Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. Nat. Biotechnol., 22, 911–917.

16. Westbrook, J., Feng, Z., Jain, S., Bhat, T.N., Thong, N., Ravichandran, V., Gilliland, G.L., Bluhm, W., Weissig, H., Greer, D.S. et al. (2002) The Protein Data Bank: unifying the archive. Nucleic Acids Res., 30, 245–248.

17. 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.

18. Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M.R., Wipat, A. et al. (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics, 20, 3045–3054.

19. Keshet, I.M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I.T., Peralta-Gil, M. and Karp, P.D. (2005) EcoCyc: a comprehensive database resource for Escherichia coli. Nucleic Acids Res., 33, D334–D337.

20. Jansen, R., Weng, S., Hitz, B., Balakrishnan, R., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hirschman, J.E. et al. (2007) Expanding protein interaction information at SGD: new pages and proteome browser. Nucleic Acids Res., 35, D468–D471.

21. McKusick, V.A. (1998) Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders, 12th edn. Johns Hopkins University Press, Baltimore.

22. Saric, J., Jensen, L.J., Ouzounova, R., Rojas, I. and Bork, P. (2006) Extraction of regulatory gene/protein networks from Medline. Bioinformatics, 22, 645–650.

23. Von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huyten, M.A. and Bork, P. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res., 33, D433–D437.

24. Shannon, P.T., Reiss, D.J., Bonneau, R. and Baliga, N.S. (2006) The Gaggle: an open-source software system for integrating bioinformatics software and data sources. BMC Bioinformatics, 7, 176.