D808–D815 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 29 November 2012*
doi:10.1093/nar/gks1094

STRING v9.1: protein-protein interaction networks, with increased coverage and integration

Andrea Franceschini[1], Damian Szklarczyk[2], Sune Frankild[2], Michael Kuhn[3], Milan Simonovic[1], Alexander Roth[1], Jianyi Lin[4], Pablo Minguez[5], Peer Bork[5,6,*], Christian von Mering[1,*] and Lars J. Jensen[2,*]

[1]Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Switzerland, [2]Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Denmark, [3]Biotechnology Center, Technical University Dresden, Germany, [4]Department of Computer Science, University of Milan, Italy, [5]European Molecular Biology Laboratory, Heidelberg and [6]Max-Delbrück-Centre for Molecular Medicine, Berlin, Germany

Received September 15, 2012; Revised October 15, 2012; Accepted October 18, 2012

ABSTRACT

Complete knowledge of all direct and indirect interactions between proteins in a given cell would represent an important milestone towards a comprehensive description of cellular mechanisms and functions. Although this goal is still elusive, considerable progress has been made—particularly for certain model organisms and functional systems. Currently, protein interactions and associations are annotated at various levels of detail in online resources, ranging from raw data repositories to highly formalized pathway databases. For many applications, a global view of all the available interaction data is desirable, including lower-quality data and/or computational predictions. The STRING database (http://string-db.org/) aims to provide such a global perspective for as many organisms as feasible. Known and predicted associations are scored and integrated, resulting in comprehensive protein networks covering >1100 organisms. Here, we describe the update to version 9.1 of STRING, introducing several improvements: (i) we extend the automated mining of scientific texts for interaction information, to now also include full-text articles; (ii) we entirely re-designed the algorithm for transferring interactions from one model organism to the other; and (iii) we provide users with statistical information on any functional enrichment observed in their networks.

INTRODUCTION

Highly complex organisms and behaviors can arise from a surprisingly restricted set of existing gene families (1,2), by a tightly regulated network of interactions among the proteins encoded by the genes. This functional web of protein–protein links extends well beyond direct physical interactions only; indeed, physical interactions might also be rather limited, covering perhaps <1% of the theoretically possible interaction space (3). Proteins do not necessarily need to undergo a stable physical interaction to have a specific, functional interplay: they can catalyze subsequent reactions in a metabolic pathway, regulate each other transcriptionally or post-transcriptionally, or jointly contribute to larger, structural assemblies without ever making direct contact. Together with direct, physical interactions, such indirect interactions constitute the larger superset of 'functional protein–protein associations' or 'functional protein linkages' (4,5).

Protein–protein associations have proven to be a useful concept, by which to group and organize all protein-coding genes in a genome. The complete set of associations can be assembled into a large network, which captures the current knowledge on the functional modularity and interconnectivity in the cell. Apart from *ad hoc* use—i.e. by browsing networks for genes of interest, inspecting interaction evidence or performing interactive clustering—a variety of systematic and large-scale usage scenarios for functional association networks have emerged. For example, (i) association networks have been frequently used to interpret the results of genome-wide genetic screens, in particular RNAi perturbation screens (6–9). Because such screens can be noisy and difficult to

*To whom correspondence should be addressed. Tel: +41 44 6353147; Fax: +41 44 6356864; Email: mering@imls.uzh.ch
Correspondence may also be addressed to Peer Bork. 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

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

interpret, any protein-network information that may help to connect potential hits can serve to provide additional confidence, particularly if a number of hits can be observed in a densely connected functional module in the network. (ii) Protein network information can aid in the interpretation of functional genomics data, e.g. in systematic proteomics surveys (10–12). This is particularly useful when the proteomics data themselves contain a protein–protein association component, such as in MS-based interaction discovery or in large-scale enzyme/substrate analysis. (iii) Protein association networks have also proven surprisingly useful for the elucidation of disease genes, both for Mendelian and for complex diseases (13–15). For the latter application, the networks can help to constrain the search space—genomic regions encompassing more than one candidate gene, or lists of genes observed to be mutated in sequencing studies, can be filtered for those genes that have connections to known disease genes (or for genes having above-random connectivity among themselves).

The STRING database has been designed with the goal to assemble, evaluate and disseminate protein–protein association information, in a user-friendly and comprehensive manner. As interactions between proteins represent such a crucial component for modern biology, STRING is by far not the only online resource dedicated to this topic. Apart from the primary databases that hold the experimental data in this field (16–20) and hand-curated databases serving expert annotations (21,22), a number of resources take a meta-analysis approach, similar to STRING. These include GeneMANIA (23), ConsensusPathDB (24), I2D (25), VisANT (26) and, more recently, hPRINT (27), HitPredict (28), IMID (29) and IMP (30). Within this wide variety of online resources and databases dedicated to interactions, STRING specializes in three ways: (i) it provides uniquely comprehensive coverage, with >1000 organisms, 5 million proteins and >200 million interactions stored; (ii) it is one of very few sites to hold experimental, predicted and transferred interactions, together with interactions obtained through text mining; and (iii) it includes a wealth of accessory information, such as protein domains and protein structures, improving its day-to-day value for users.

We have already discussed many aspects of the STRING resource previously, e.g. (31,32), including its data-sources, prediction algorithms and user-interface. Here, we describe the current update to version 9.1 of the resource, focusing on new features and updated algorithms. In particular, we will describe how STRING increasingly makes use of externally provided orthology information [from the eggNOG database (33)] to better integrate evidence across distinct organisms.

UPDATED TEXT MINING

The new version of STRING features a redesigned text-mining pipeline. We have improved the named entity recognition engine to use custom-made hashing and string-compare functions to comprehensively and efficiently handle orthographic variation related to whether a name is written as one word, two words or with a hyphen. As in the previous versions of STRING, associations between proteins are derived from statistical analysis of co-occurrence in documents and from natural language processing. The latter combines part-of-speech tagging, semantic tagging and a chunking grammar to achieve rule-based extraction of physical and regulatory interactions, as described previously (34).

To improve the quality and number of links derived from co-occurrence, we have developed an entirely new scoring scheme, which takes into account co-occurrences within sentences, within paragraphs and within whole documents and combines them through an optimized weighting scheme.

The scoring scheme first calculates a weighted count (C_{ij}) for each pair of entities i and j:

$$C_{ij} = \sum_{k=1}^{n} \delta_{dijk} w_d + \delta_{pijk} w_p + \delta_{sijk} w_s$$

where $w_d = 1$, $w_p = 2$ and $w_s = 0.2$ are the weights for co-occurrence within the same document, same paragraph and same sentence, respectively. The delta functions δ_{dijk}, δ_{pijk} and δ_{sijk} are 1, if the entities i and j are co-mentioned in the document k, a paragraph of k or a sentence of k. Based on the weighted counts, the co-occurrence score (S_{ij}) is defined as:

$$S_{ij} = C_{ij}^{\alpha} \left(\frac{C_{ij} C_{\bullet\bullet}}{C_{i\bullet} C_{\bullet j}} \right)^{1-\alpha}$$

where $C_{\text{i}\bullet}$ and $C_{\bullet\text{j}}$ are the sums over all pairs involving i or j and an entity from the same taxon, $C_{\bullet\bullet}$ is the sum over all pairs of entities from the taxon, and $\alpha = 0.6$. The parameters were optimized on the KEGG benchmark set.

This has substantially improved the quality and number of associations extracted (Table 1). The more efficient named entity recognition engine and the new scoring scheme also enabled us to move beyond the parsing of MEDLINE abstracts, and to now include text mining of 1 821 983 full-text articles, which were freely available from publishers web sites. This has further improved the comprehensiveness of the text mining in the new version of STRING (Table 1). The natural language processing part of the pipeline has also been standardized, to make use of an ontology that describes possible molecular modes of action by which proteins can influence each other (35). Finally, the new text-mining pipeline explicitly takes into account orthology information by treating each orthologous group as an entity that is considered whenever one of its member proteins is mentioned (33), thereby directly detecting associations between orthologous groups as well as between proteins.

TRANSFER OF INTERACTIONS BETWEEN ORGANISMS

Evolutionarily related proteins are known to usually maintain their three-dimensional structure, even when they have become so diverged over time that there is hardly any detectable sequence similarity left between them

Table 1. Protein–protein associations based on automated text mining

	STRING v9.0	STRING v9.1	Fold increase
Natural language processing	38 859	63 331	1.629
Cooccurrence, high confidence	286 880	792 730	2.763
Cooccurrence, medium confidence	1 100 756	1 672 222	1.519
Cooccurrence, low confidence	3 214 754	4 270 322	1.328

This table quantifies non-redundant associations extracted by text mining in STRING, at various confidence levels; note that both STRING versions shown here are based on the same set of organisms and proteins. The increase in text-mining interactions is largest in the high confidence bracket, reflecting the increased performance enabled by the extension to full text articles, and by the improved entity recognition engine.

(36,37). Similarly, most protein–protein interaction interfaces remain well-conserved over time, at least for the case of stably bound protein partners located next to each other in protein complexes (38,39). This means that a pair of proteins observed to be stably binding in one organism can be expected to be binding in another organism as well, provided both genes have been retained in both genomes. The term ‘interologs’ was coined for such pairs, a combination of the words ‘interaction’ and ‘ortholog’ (40). Whether this high degree of interaction conservation is true also for other, more indirect or transient types of protein–protein associations is less clear—although at least one such type, namely joint metabolic pathway membership, has also been shown to be generally well-conserved (41,42). Based on the principle of interaction conservation, evidence transfer from one model organism to the other seems feasible, and it has been implemented in several frameworks already.

In practice, the search for potential interologs is not trivial, except for very closely related organisms. The reason for this lies in the high frequency of gene duplications, gene losses and gene re-arrangements, which makes it difficult to assign pairs of functionally equivalent genes across distant organisms. The best candidates for functionally equivalent genes in two organisms are ‘one-to-one’ orthologs, i.e. genes that track back to a single gene in the last common ancestor of both organisms, and have since undergone little or no duplication or loss events (43–45). In a large resource such as STRING, unequivocally identifying one-to-one orthologs for all pairs of organisms is not feasible: there are potentially more than a million pairs of organisms to study, each with thousands of genes, and the proper identification of orthologs would ideally entail exhaustive and time-consuming phylogenetic tree analysis. In the past, STRING has therefore used two distinct heuristic options: either to substitute homology for orthology (46) or to use pre-defined orthology relations described at high-level taxonomic groups, from the COG database (47). We found that both approaches were suboptimal; they both transferred evidence even when the presence of multiple paralogs indicated that the orthology situation was somewhat unclear—despite an explicit procedure to down-weigh the transferred scores in such cases, at least in the homology approach (46). We have, therefore, now devised a procedure that more explicitly considers the known phylogeny of organisms and which works on the basis of hierarchical orthologous groups maintained at the eggNOG database (33).

The taxonomy tree covering the 1133 species present in STRING consists of 495 branching nodes at different taxonomic positions (the tree is a down-sampled version of the taxonomy maintained at NCBI). Through experimentation and benchmarking, we have developed a new two-step procedure, which makes use of this tree for the transfer of functional associations. First, associations between proteins are transferred to the orthologous groups to which the proteins belong; this proceeds sequentially from lower to increasingly higher levels of taxonomic hierarchy. Second, associations are transferred in the opposite direction, i.e. from the orthologous groups back to their constituent proteins. Where available, the hierarchical orthology groups from eggNOG version 3 are used (33). As many of the taxonomic positions in the tree are not covered in eggNOG, we construct provisional groups for the missing positions by down-sampling the orthologous groups from the next higher taxonomy level present in eggNOG.

To compute a score of functional association (S_{abk}) between two orthologous groups a and b at the taxonomic level k, we sort the n associations (P_{abi}) between their member proteins from highest to lowest score, and then integrate them sequentially (Figure 1):

$$S_{abk} = 1 - (1 - p')\left(\prod_{i=1}^{n} \frac{1 - P_{abi} f_{abi}^{\alpha} \min_{j} d_{ij}}{1 - p'}\right)$$

where p' is prior probability of two proteins being linked, which is 0.063 according to the KEGG benchmark set; f_{abi} is a penalty dependent on the number of paralogs of a given protein pair and d_{ij} is a penalty dependent on the similarity of the species i and the other species j that have already been included in the score:

$$f_{abi} = \left(\frac{1}{c_{ai} c_{bi}}\right) \quad d_{ij} = 1 - \frac{1}{1 + \exp[\beta(\delta - s_{ij})]}$$

where c_{ai} and c_{bi} are the number of proteins from a given species in the orthologous groups, and s_{ij} the median similarity between the given species, measured on a universal set of marker gene families (48) and expressed as the ‘self-normalized bit-score’ (i.e. the bit score of an alignment between two proteins, which is divided by the bit score of a self-alignment of the shorter of the two proteins; this measure always ranges from zero to one).

The process is repeated for all pairs of orthologous groups at every taxonomic level. Next, the scores between pairs of orthologous groups are transferred

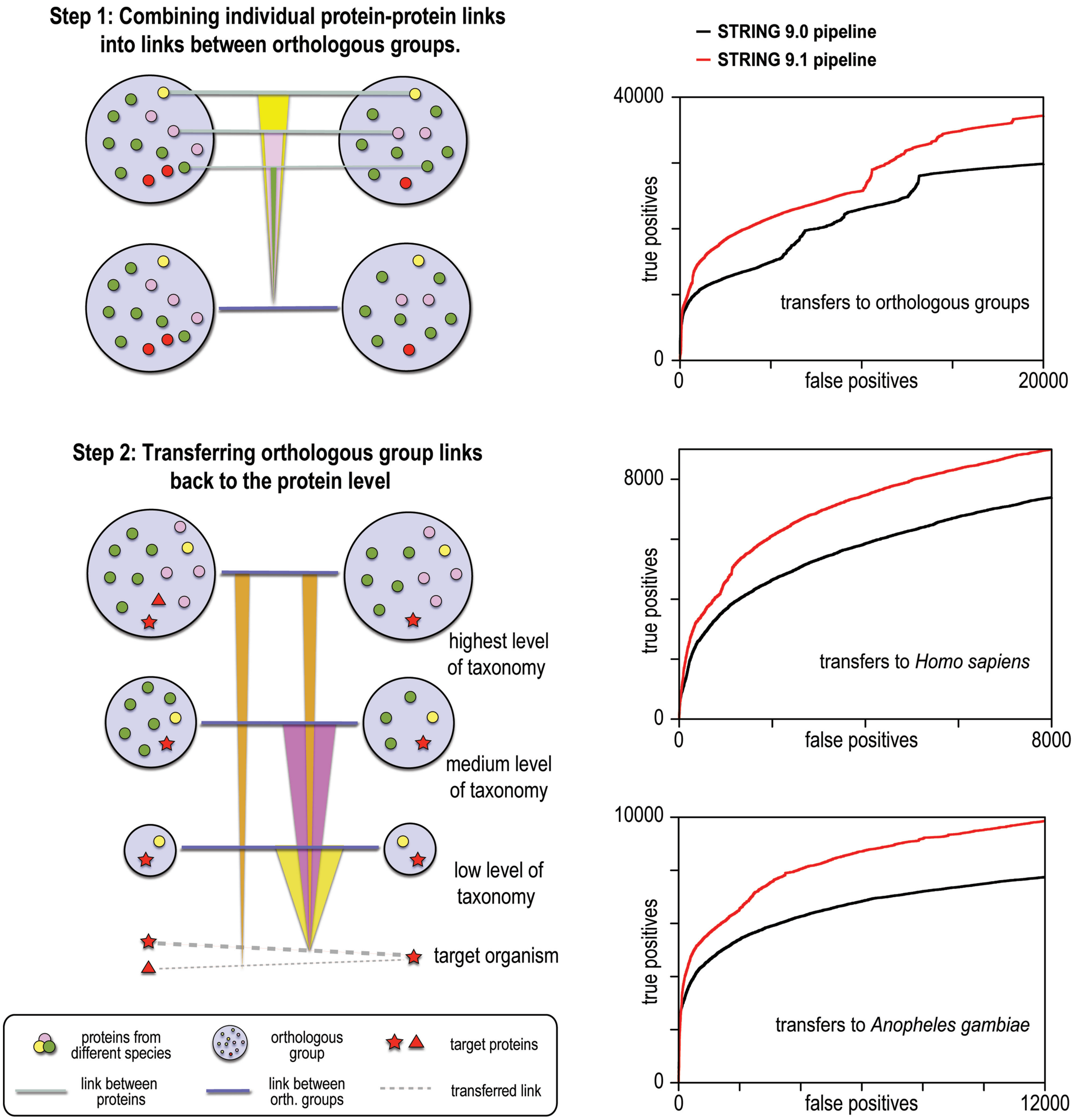


Figure 1. Improved procedure for interaction transfer between organisms. Left: steps 1 and 2 of the functional association transfer pipeline. In the first step, the individual links between proteins are combined into a score between orthologous groups, sequentially, from the strongest link (thick line) to the weakest (thin). Each subsequent score is down-weighted, both based on the similarity of its organism to organisms that have already contributed to the combined scores, and on number of proteins from the same organism inside the orthologous group. In the second step of the transfer pipeline, the links between orthologous groups are transferred back to individual protein pairs belonging to these groups. This is done sequentially from the lowest to highest taxonomy level. In the above example, the two transferred links from the highest taxonomic level (orange links) are penalized for the increase in number of proteins from the target species in one of the orthologous groups. Right: ROC curves indicating the performance of predicted interolog scores, benchmarked against KEGG pathways; an inferred link between two proteins is considered to be a true positive when both proteins are annotated to be together in at least one shared KEGG pathway.

back to protein pairs; this finally results in the actual evidence transfer between organisms. To calculate the transferred score (T_{im}) from all taxonomic levels m to a protein pair from species i, we combine the scores (S_{abk}) from orthologous groups consecutively from the lowest to the highest taxonomy level, subtracting the contributions from all lower taxonomic levels (Figure 1):

$$T_{im} = 1 - (1 - p') \prod_{k=1}^{m} \frac{1 - S_{abk} f_{abi}^{\varepsilon} \min(s_a,s_b)^{\gamma}}{(1 - T_{i,k-1})(1 - P_{abi})(1 - p')}$$

where at each taxonomic level, we subtract the part of the score that originates from the species itself (P_{abi}) while additionally penalizing it for the number of paralogs in the respective orthologous groups (f_{abi}) and for the median self-normalized bit scores (s_a and s_b) of the proteins in the groups a and b.

The parameters α, ε and γ are universal in the sense that they have the same values for all evidence channels in STRING, e.g. co-occurence, experiments and text mining, whereas β and δ are channel specific to take into account the different rate at which scores become independent from each other. The new transfer scheme was optimized and benchmarked on the set of known interactions in the KEGG database and achieves better performance than the previous method, both for orthologous groups and for individual proteins (Figure 1).

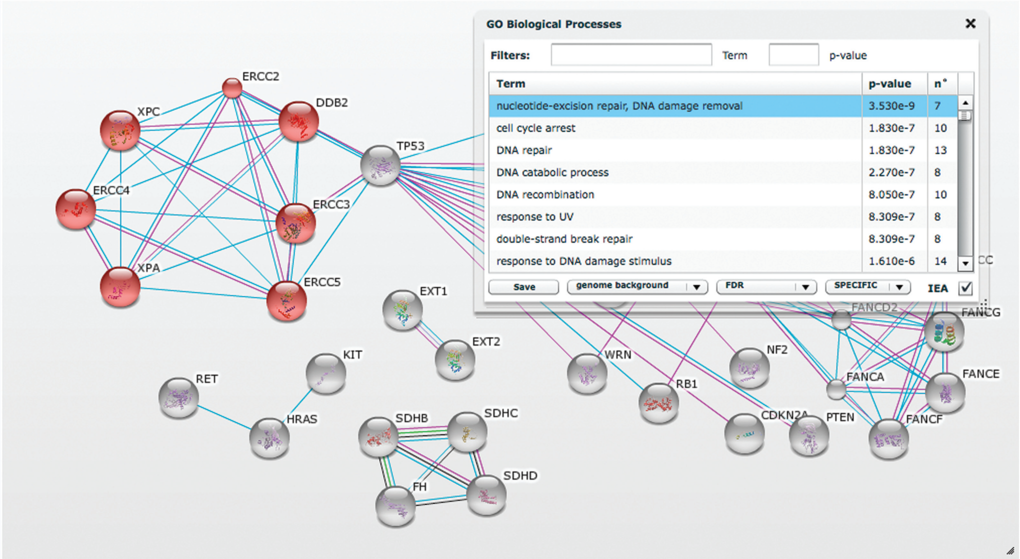
STATISTICAL ENRICHMENT ANALYSIS

STRING users that do not just query with a single protein of interest, but instead upload entire lists of proteins, are often interested in knowing whether their input shows evidence for a statistical enrichment of any known biological function or pathway. To address this question, a variety of dedicated online resources are already available (49,50), most notably the DAVID resource (51). However, entering gene lists at multiple websites can be cumbersome, and not all existing resources will make full use of the latest protein network information. Therefore, we have now included functionality to detect enrichment of functional systems in each currently displayed network in STRING, testing a number of functional annotation spaces including Gene Ontology, KEGG, Pfam and InterPro (see Figure 2). Any detected enrichments can be browsed interactively, visually highlighting the corresponding proteins in the network (Figure 2).

In the Enrichment widget, STRING displays every functional pathway/term that can be associated to at least one protein in the network. The terms are sorted by their enrichment *P*-value, which we compute using a Hypergeometric test, as explained in (53). The *P*-values are corrected for multiple testing using the method of Benjamini and Hochberg (54), but we also provide options to either disable that correction or to select a more stringent statistical test (Bonferroni). In the case of testing for Gene Ontology enrichments, users have the additional options to exclude annotations inferred by automatic procedures only (Electronic Inferred Associations), to limit the testing to pre-defined higher level categories (GO Slim), or to prune away parent terms that are redundant with child terms (i.e. covering the exact same set of proteins).

Furthermore, we report to the user whether the protein list is enriched in STRING interactions *per se*, independent of known pathway annotations. The latter functionality is non-trivial and requires an explicit null model, owing to the non-uniform distribution of the connectivity degrees of proteins in networks (9,55–57). We chose a random background model that preserves the degree distribution of the proteins in a given list: the Random Graph with Given Degree Sequence (RGGDS), similar to references (55,57).

Given a list L of proteins, let X_L denote the number of edges connecting proteins in an RGGDS with similar size as L. For the given $L_,$ a strong edge enrichment



GO Biological Processes ×

Filters: Term p-value

Term	p-value	n°
nucleotide-excision repair, DNA damage removal	3.530e-9	7
cell cycle arrest	1.830e-7	10
DNA repair	1.830e-7	13
DNA catabolic process	2.270e-7	8
DNA recombination	8.050e-7	10
response to UV	8.309e-7	8
double-strand break repair	8.309e-7	8
response to DNA damage stimulus	1.610e-6	14

Save | genome background ▾ | FDR ▾ | SPECIFIC ▾ | IEA ☑

Figure 2. Network visualization and statistical analysis of a user-supplied protein list. The STRING screenshot shows a user-supplied set of genes, here a selection of cancer genes as annotated at the COSMIC database (52). The set is restricted to those genes that are known to pre-dispose to cancer already when mutated in the germline, and that have at least one connection in STRING. The inset illustrates the website's new functionality for automatically detecting statistically enriched functions or processes in a network. In this example, one of the detected processes (nucleotide excision repair) is of interest and has been selected; STRING automatically highlighted the corresponding nodes in the network, where they are seen to form a densely connected module.

corresponds to a low probability of counting, in the RGGDS, at least the observed number x of edges connecting proteins in L, i.e. a low value of:

$$S_L(x) = P(X_L \geq x)$$

The random variable X_L is a sum of Bernoulli variables with distinct parameters, and hence a Poisson–Binomial variable. If L is large, X_L can thus be approximated by a Poisson random variable, whose cumulative probability function is:

$$S_L(x) = P(X_L \geq x) \cong \frac{1}{\alpha}\sum_{n=x}^{M}\frac{e^{-\lambda}\lambda^n}{n!},$$

$$\lambda = \frac{1}{2}\sum_{\substack{u,v\in L\\ u\neq v}} P_{uv},$$

$$\alpha = \sum_{n=0}^{M}\frac{e^{-\lambda}\lambda^n}{n!}, \quad p_{ij} \cong 1 - \exp\left(-\frac{\deg(v_i)\deg(v_i)}{2M}\right)$$

with M being the total number of interactions within L in STRING, and *deg(v)* denoting the degree of protein v, i.e. the number of interaction partners it has.

USER INTERFACE

The STRING website aims to provide easy and intuitive interfaces for searching and browsing the protein interaction data, as well as for inspecting the underlying evidence. Users can query for a single protein of interest, or for a set of proteins, using a variety of different identifier name spaces. The resulting network can then be inspected, rearranged interactively or clustered at variable stringency. Each protein node in the network shows a preview to 3D structural information, if available, and can be clicked to reveal a pop-up window with more information about the protein [including its annotation (58), SMART domain-structure (59), structure homology models from SWISS-MODEL Repository (60), etc.]. Each edge in the network denotes a known or predicted interaction, and leads to a pop-up window providing details on the underlying evidence and the interaction confidence scores.

An important new feature in version 9.1 of STRING is the possibility for users to identify themselves by logging in. Although this is not necessary for basic browsing and searching, it provides users with the option to browse their history of past searches, save visited pages for later return and upload lists of proteins that are of interest to them. In addition, logging in is useful for storing and retrieving 'payload' information to be shown and browsed alongside the network. As described previously (31), 'payload' information is user-provided extra data that can be projected onto the STRING network; it can consist of information regarding both nodes (proteins) and edges (interactions). Previously, any payload information had to be communicated to STRING via a set of files following a specific format—now, they can be uploaded and managed interactively.

ACKNOWLEDGEMENTS

The authors wish to thank Yan P. Yuan (EMBL) for excellent administrative support with the STRING backend servers, and Carlos García Girón (Sanger Institute) for help in implementing the user-payload-data mechanism.

FUNDING

The Swiss Institute of Bioinformatics (SIB) provides sustained funding for this project. Work on the project has also been supported in part by the Novo Nordisk Foundation Center for Protein Research and the European Molecular Biology Laboratory (EMBL). Funding for open access charge: University of Zurich.

Conflict of interest statement. None declared.

REFERENCES

1. Chothia,C. (1992) Proteins. One thousand families for the molecular biologist. *Nature*, **357**, 543–544.
2. Wolf,Y.I., Grishin,N.V. and Koonin,E.V. (2000) Estimating the number of protein folds and families from complete genome data. *J.Mol. Biol.*, **299**, 897–905.
3. Aloy,P. and Russell,R.B. (2004) Ten thousand interactions for the molecular biologist. *Nature Biotechnol.*, **22**, 1317–1321.
4. Huynen,M., Snel,B., Lathe,W. 3rd and Bork,P. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.*, **10**, 1204–1210.
5. Eisenberg,D., Marcotte,E.M., Xenarios,I. and Yeates,T.O. (2000) Protein function in the post-genomic era. *Nature*, **405**, 823–826.
6. Gonzalez,O. and Zimmer,R. (2011) Contextual analysis of RNAi-based functional screens using interaction networks. *Bioinformatics*, **27**, 2707–2713.
7. Simpson,J.C., Joggerst,B., Laketa,V., Verissimo,F., Cetin,C., Erfle,H., Bexiga,M.G., Singan,V.R., Heriche,J.K., Neumann,B. *et al.* (2012) Genome-wide RNAi screening identifies human proteins with a regulatory function in the early secretory pathway. *Nature Cell Biol.*, **14**, 764–774.
8. Moreau,D., Kumar,P., Wang,S.C., Chaumet,A., Chew,S.Y., Chevalley,H. and Bard,F. (2011) Genome-wide RNAi screens identify genes required for Ricin and PE intoxications. *Dev. Cell*, **21**, 231–244.
9. Kaplow,I.M., Singh,R., Friedman,A., Bakal,C., Perrimon,N. and Berger,B. (2009) RNAiCut: automated detection of significant genes from functional genomic screens. *Nat. Methods*, **6**, 476–477.
10. Goh,W.W., Lee,Y.H., Chung,M. and Wong,L. (2012) How advancement in biological network analysis methods empowers proteomics. *Proteomics*, **12**, 550–563.
11. Oppermann,F.S., Grundner-Culemann,K., Kumar,C., Gruss,O.J., Jallepalli,P.V. and Daub,H. (2012) Combination of chemical genetics and phosphoproteomics for kinase signaling analysis enables confident identification of cellular downstream targets. *Mol. Cell. Proteomics*, **11**, O111 012351.
12. Olsson,N., James,P., Borrebaeck,C.A. and Wingren,C. (2012) Quantitative proteomics targeting classes of motif-containing peptides using immunoaffinity-based mass spectrometry. *Mol. Cell. Proteomics*, **11**, 342–354.
13. Lee,I., Blom,U.M., Wang,P.I., Shim,J.E. and Marcotte,M. (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.*, **21**, 1109–1121.
14. Moreau,Y. and Tranchevent,L.C. (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat. Rev. Genet.*, **13**, 523–536.

15. Piro,R.M. and Di Cunto,F. (2012) Computational approaches to disease-gene prediction: rationale, classification and successes. *FEBS J.*, **279**, 678–696.
16. Stark,C., Breitkreutz,B.J., Chatr-Aryamontri,A., Boucher,L., Oughtred,R., Livstone,M.S., Nixon,J., Van Auken,K., Wang,X., Shi,X. *et al.* (2011) The BioGRID interaction database: 2011 update. *Nucleic Acids Res.*, **39**, D698–D704.
17. Kerrien,S., Aranda,B., Breuza,L., Bridge,A., Broackes-Carter,F., Chen,C., Duesbury,M., Dumousseau,M., Feuermann,M., Hinz,U. *et al.* (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, **40**, D841–D846.
18. 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.
19. Licata,L., Briganti,L., Peluso,D., Perfetto,L., Iannuccelli,M., Galeota,E., Sacco,F., Palma,A., Nardozza,A.P., Santonico,E. *et al.* (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.*, **40**, D857–D861.
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. Goel,R., Harsha,H.C., Pandey,A. and Prasad,T.S. (2012) Human protein reference database and human proteinpedia as resources for phosphoproteome analysis. *Mol. Biosyst.*, **8**, 453–463.
22. Croft,D., O'Kelly,G., Wu,G., Haw,R., Gillespie,M., Matthews,L., Caudy,M., Garapati,P., Gopinath,G., Jassal,B. *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691–D697.
23. 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**, W214–W220.
24. Kamburov,A., Pentchev,K., Galicka,H., Wierling,C., Lehrach,H. and Herwig,R. (2011) ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res.*, **39**, D712–D717.
25. 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.
26. 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.
27. Elefsinioti,A., Sarac,O.S., Hegele,A., Plake,C., Hubner,N.C., Poser,I., Sarov,M., Hyman,A., Mann,M., Schroeder,M. *et al.* (2011) Large-scale de novo prediction of physical protein-protein association. *Mol. Cell. Proteomics*, **10**, M111 010629.
28. Patil,A., Nakai,K. and Nakamura,H. (2011) HitPredict: a database of quality assessed protein-protein interactions in nine species. *Nucleic Acids Res.*, **39**, D744–D749.
29. Balaji,S., McClendon,C., Chowdhary,R., Liu,J.S. and Zhang,J. (2012) IMID: integrated molecular interaction database. *Bioinformatics*, **28**, 747–749.
30. Wong,A.K., Park,C.Y., Greene,C.S., Bongo,L.A., Guan,Y. and Troyanskaya,O.G. (2012) IMP: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. *Nucleic Acids Res.*, **40**, W484–W490.
31. Szklarczyk,D., Franceschini,A., Kuhn,M., Simonovic,M., Roth,A., Minguez,P., Doerks,T., Stark,M., Muller,J., Bork,P. *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, **39**, D561–D568.
32. 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.
33. Powell,S., Szklarczyk,D., Trachana,K., Roth,A., Kuhn,M., Muller,J., Arnold,R., Rattei,T., Letunic,I., Doerks,T. *et al.* (2012) eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.*, **40**, D284–D289.
34. 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.
35. Minguez,P., Parca,L., Diella,F., Mende,D.R., Kumar,R., Helmer-Citterich,M., Gavin,A.C., van Noort,V. and Bork,P. (2012) Deciphering a global network of functionally associated post-translational modifications. *Mol. Syst. Biol.*, **8**, 599.
36. Thornton,J.M., Orengo,C.A., Todd,A.E. and Pearl,F.M. (1999) Protein folds, functions and evolution. *J. Mol. Biol.*, **293**, 333–342.
37. Koonin,E.V., Wolf,Y.I. and Karev,G.P. (2002) The structure of the protein universe and genome evolution. *Nature*, **420**, 218–223.
38. Zhang,Q.C., Petrey,D., Norel,R. and Honig,B.H. (2010) Protein interface conservation across structure space. *Proc. Natl Acad. Sci. USA*, **107**, 10896–10901.
39. Qian,W., He,X., Chan,E., Xu,H. and Zhang,J. (2011) Measuring the evolutionary rate of protein-protein interaction. *Proc. Natl Acad. Sci. USA*, **108**, 8725–8730.
40. Walhout,A.J., Sordella,R., Lu,X., Hartley,J.L., Temple,G.F., Brasch,M.A., Thierry-Mieg,N. and Vidal,M. (2000) Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science*, **287**, 116–122.
41. Caspi,R., Foerster,H., Fulcher,C.A., Kaipa,P., Krummenacker,M., Latendresse,M., Paley,S., Rhee,S.Y., Shearer,A.G., Tissier,C. *et al.* (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, **36**, D623–D631.
42. Teichmann,S.A., Rison,S.C., Thornton,J.M., Riley,M., Gough,J. and Chothia,C. (2001) The evolution and structural anatomy of the small molecule metabolic pathways in Escherichia coli. *J. Mol. Biol.*, **311**, 693–708.
43. Conant,G.C. and Wolfe,K.H. (2008) Turning a hobby into a job: how duplicated genes find new functions. *Nat. Rev. Genet.*, **9**, 938–950.
44. Koonin,E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Ann. Rev. Genet.*, **39**, 309–338.
45. Altenhoff,A.M., Studer,R.A., Robinson-Rechavi,M. and Dessimoz,C. (2012) Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comput. Biol.*, **8**, e1002514.
46. von Mering,C., Jensen,L.J., Snel,B., Hooper,S.D., Krupp,M., Foglierini,M., Jouffre,N., Huynen,M.A. and Bork,P. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.*, **33**, D433–D437.
47. Tatusov,R.L., Galperin,M.Y., Natale,D.A. and Koonin,E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33–36.
48. Ciccarelli,F.D., Doerks,T., von Mering,C., Creevey,C.J., Snel,B. and Bork,P. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science*, **311**, 1283–1287.
49. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
50. Khatri,P., Sirota,M. and Butte,A.J. (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.*, **8**, e1002375.
51. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
52. Forbes,S.A., Bindal,N., Bamford,S., Cole,C., Kok,C.Y., Beare,D., Jia,M., Shepherd,R., Leung,K., Menzies,A. *et al.* (2011) COSMIC: mining complete cancer genomes in the Catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, **39**, D945–D950.
53. Rivals,I., Personnaz,L., Taing,L. and Potier,M.C. (2007) Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, **23**, 401–407.

54. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. B*, **57**, 289–300.
55. Maslov,S. and Sneppen,K. (2002) Specificity and stability in topology of protein networks. *Science*, **296**, 910–913.
56. Minguez,P., Gotz,S., Montaner,D., Al-Shahrour,F. and Dopazo,J. (2009) SNOW, a web-based tool for the statistical analysis of protein-protein interaction networks. *Nucleic Acids Res.*, **37**, W109–W114.
57. Pradines,J.R., Farutin,V., Rowley,S. and Dancik,V. (2005) Analyzing protein lists with large networks: edge-count probabilities in random graphs with given expected degrees. *J. Comput. Biol.*, **12**, 113–128.
58. Apweiler,R., Martin,M.J., O'Donovan,C., Magrane,M., Alam-Faruque,Y., Antunes,R., Barrell,D., Bely,B., Bingley,M., Binns,D. *et al.* (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, **39**, D214–D219.
59. Letunic,I., Doerks,T. and Bork,P. (2012) SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res.*, **40**, D302–D305.
60. 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.