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

Andrea Franceschini1, Damian Szklarczyk2, Sune Frankild2, Michael Kuhn3, Milan Simonovic1, Alexander Roth1, Jianyi Lin4, Pablo Minguez5, Peer Bork5,6,*; Christian von Mering1,* and Lars J. Jensen2,*

1Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Switzerland, 2Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Denmark, 3Biotechnology Center, Technical University Dresden, Germany, 4Department of Computer Science, University of Milan, Italy, 5European Molecular Biology Laboratory, Heidelberg and 6Max-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...
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 \( j \):

\[
C_{ij} = \sum_{k=1}^{n} \delta_{dijk} w_d + \delta_{pijk} w_p + \delta_{ijk} w_s
\]

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 \( \delta_{dijk} \), \( \delta_{pijk} \) and \( \delta_{ijk} \) 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:

\[
S_{ij} = C_{ij} \left( \frac{C_{ij} C_{**}}{C_{**}} \right)^{1-\alpha}
\]

where \( C_{**} \) is the sum over all pairs involving \( i \) or \( j \) and an entity from the same taxon, \( C_{**} \) 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.
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):

\[
S_{abk} = 1 - (1 - p') \left( \prod_{i=1}^{n} \frac{1 - P_{abi} f_{abi}^* \min(d_{ij})}{1 - p'} \right)
\]

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:

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

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...
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):

\[
T_{im} = 1 - (1 - p') \prod_{k=1}^{m} \frac{1 - S_{abk}f_{abk}}{1 - T_{ik-1}f_{abk}} \min(s_{ab}, s_{b})^{p'}
\]
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 \(\alpha, \varepsilon\) and \(\gamma\) are universal in the sense that they have the same values for all evidence channels in STRING, e.g. co-occurrence, experiments and text mining, whereas \(\beta\) and \(\delta\) 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

![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.](image-url)
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:

\[
S_L(x) = P(X_L \geq x) \approx \frac{1}{M} \sum_{n=0}^{M} \frac{e^{-\lambda} \lambda^n}{n!},
\]

\[
\lambda = \sum_{u,v \in L} P_{uv},
\]

\[
\alpha = \sum_{v \in L} \frac{e^{-\alpha} \alpha^{deg(v)}}{n!}, \quad p_{uv} \equiv 1 - \exp\left(-\frac{deg(v_i) \cdot deg(v_j)}{2M}\right)
\]

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 a 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**


