STRING: known and predicted protein–protein associations, integrated and transferred across organisms

Christian von Mering, Lars J. Jensen, Berend Snel, Sean D. Hooper, Markus Krupp, Mathilde Foglierini, Nelly Jouffre, Martijn A. Huynen and Peer Bork*

European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany and \(^1\)Nijmegen Centre for Molecular Life Sciences / Centre of Molecular and Biomolecular Informatics, University Medical Center St Radboud, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands

Received September 11, 2004; Accepted September 13, 2004

ABSTRACT

A full description of a protein’s function requires knowledge of all partner proteins with which it specifically associates. From a functional perspective, ‘association’ can mean direct physical binding, but can also mean indirect interaction such as participation in the same metabolic pathway or cellular process. Currently, information about protein association is scattered over a wide variety of resources and model organisms. STRING aims to simplify access to this information by providing a comprehensive, yet quality-controlled collection of protein–protein associations for a large number of organisms. The associations are derived from high-throughput experimental data, from the mining of databases and literature, and from predictions based on genomic context analysis. STRING integrates and ranks these associations by benchmarking them against a common reference set, and presents evidence in a consistent and intuitive web interface. Importantly, the associations are extended beyond the organism in which they were originally described, by automatic transfer to orthologous protein pairs in other organisms, where applicable. STRING currently holds 730 000 proteins in 180 fully sequenced organisms, and is available at http://string.embl.de/.

INTRODUCTION

Several databases exist, whose main purpose is to collect and curate direct experimental evidence about protein–protein interactions (1–4). Other databases take a more generalized perspective on proteins and their associations, by functionally grouping proteins into metabolic, signaling or transcriptional pathways (5–8). Finally, a third class of resources attempts to fill gaps in both datasets, by predicting protein–protein associations de novo, using a variety of computational techniques (9–13).

The database STRING (‘Search Tool for the Retrieval of Interacting Genes/Proteins’) represents an ongoing effort to provide these three types of protein–protein association evidence under one common framework. Such an integrated approach offers several unique advantages: (i) various types of evidence are mapped onto a single, stable set of proteins, thereby facilitating comparative analysis; (ii) known and predicted interactions often partially complement each other, leading to increased coverage; (iii) an integrated scoring scheme can provide higher confidence when independent evidence types agree; and (iv) mapping and transferring interactions onto a large number of organisms facilitates evolutionary studies.

Because STRING is fully pre-computed, all information can be quickly accessed—both at the high-level network view and at the level of the individual interaction record. The various evidence types can be enabled or disabled separately, which allows the searches to be customized at run-time, and dedicated viewers allow the inspection of all the evidence underlying an association (Figure 1). The database is an exploratory resource: it contains a much larger number of associations than primary interaction databases—albeit with varying confidence scores. It is thus best used for getting a quick initial overview of the functional partners of a query protein, especially for proteins that are still poorly characterized.

DATA SOURCES AND SCORING

Many of the protein–protein associations in STRING are imported from other databases (see below), but STRING

*To whom correspondence should be addressed. Tel: +49 6221 387 8526; Fax: +49 6221 387 517; Email: mering@embl-heidelberg.de

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use permissions, please contact journals.permissions@oupjournals.org.

© 2005, the authors
also contains a large body of predicted associations that are produced de novo. These predictions are based on systematic genome comparisons ['genomic context', (14,15)]. We periodically import completely sequenced genomes [metazoan genomes from Ensembl, all others from SwissProt, (16)], and search them for three types of genomic context associations: conserved genomic neighborhood, gene fusion events, and co-occurrence of genes across genomes. All three searches aim to identify pairs of genes which appear to be under common selective pressures during evolution (more so than expected by chance), and which are therefore thought to be functionally associated.

As for all other types of associations in STRING, we assign a confidence score to each predicted association. The scores are derived by benchmarking the performance of the predictions against a common reference set of trusted, true associations. We chose the functional grouping of proteins maintained at KEGG [Kyoto Encyclopedia of Genes and Genomes, (5)] as the reference. Any predicted association for which both proteins are assigned to the same ‘KEGG pathway’ is counted as a true positive. KEGG pathways are particularly suitable as a reference because they are based on manual curation, are available for a number of organisms, and cover several functional areas. The benchmarked confidence scores in STRING generally correspond to the probability of finding the linked proteins within the same KEGG pathway. STRING performs a similar benchmark for high-throughput experimental interaction data, separately for each dataset. Scores vary within one dataset because they include additional, intrinsic information from the data itself, such as the frequency or reciprocity of the detection (see Figure 2 for a typical benchmark). In contrast to high-throughput data, validated small-scale interactions, protein complexes, and annotated pathways...
are directly imported from databases (2,5,17), and given a uniform confidence score per dataset.

Another important source of protein association information is the published literature (18,19). We systematically extract associations from PubMed, by searching for recurrent co-mentioning of gene names in abstracts. This search relies on gene names and synonyms parsed from SwissProt as well as from organism-specific databases, and we utilize a benchmarked scoring system based on the frequencies and distributions of gene names in abstracts (not shown).

Finally, we also derive protein–protein associations from functional genomics data: co-regulation of genes across diverse experimental conditions, as measured by using microarray analysis, can be a predictor of functional associations (20). We import these associations from the ArrayProspector server (12), which is based on the same benchmarks and genomes as STRING itself.

**TRANSFER OF ASSOCIATIONS ACROSS ORGANISMS**

STRING employs two different strategies for transferring known and predicted associations between organisms (Figure 3): the first (‘COG-mode’) relies on externally provided orthology assignments and transfers interactions in an all-or-none fashion, whereas the second (‘protein-mode’) uses quantitative sequence similarity searches and often distributes a given interaction fractionally among several protein pairs of the target organism. Both approaches have strengths and weaknesses, and users can choose either one of them before starting their query (a color change helps them to distinguish the modus throughout the user interface).

The COG mode requires an assignment of proteins into orthologous groups; all proteins within such a group are assumed to be functionally equivalent across genomes. This orthology information is imported from the COGs database [(21), we extend the groups to cover all organisms in STRING]. Any association score observed between a pair of proteins from different COGs is assumed to be valid for all protein pairs spanning these two COGs. Repeated observations of links, e.g. occurrence of genes in the same operon, increase the association score—but only when they are observed in phylogenetically distant organisms.

In the newly developed protein mode, there is no pre-signed orthology information. Instead, the transfer relies on a precomputed all-against-all similarity search of the 730,000 proteins in STRING (using the sensitive Smith-Waterman algorithm). For each association to be transferred, the algorithm searches for potential orthologs of the interacting partners in other genomes. Orthology is assumed if proteins form reciprocal best matches in the searches, in the absence of any close, second-best hits (paralogs) in either species. In such an ideal situation, the interactions can be transferred in toto. However, in reality there will often be additional paralogs in one or both of the genomes, which complicates the transfer. We have devised and benchmarked an empirical scheme that is based on the relative sequence similarity of competing paralogous proteins (Figure 3). Essentially, the pair of proteins exhibiting the highest sequence similarity to the source pair receives the highest ‘share’ of the transferred interaction.

**INTEGRATION**

After assignment of association scores and transfer between species, we compute a final ‘combined score’ between any pair of proteins (or pair of COGs). This score is often higher than the individual sub-scores, expressing increased confidence when an association is supported by several types of evidence (Table 1). It is computed under the assumption of independence for the various sources, in a naïve Bayesian fashion. It is thus a simple expression of the individual scores:

\[ S = 1 - \prod_i (1 - S_i) \]

The assumption of independence is valid here because datasets that are based on similar technologies (e.g. different yeast two-hybrid datasets) have been joined previously and are benchmarked as a single information source. Along with the combined score, the individual sub-scores are always displayed as well, because they provide valuable information about the nature of a particular association.
Figure 3. Transferring association scores between organisms. Initial situation (top): a scored association between two proteins in a source organism—how confidently can it be transferred to a target organism by a postulated association among homologous proteins? Bottom left: in ‘COG-mode’, all proteins in an orthologous group (COG) are considered equivalent. The highest association score between any two proteins in the two COGs is assumed to be valid for all pairs. Bottom right: in ‘protein-mode’, all sequence similarity relations between the two organisms are considered. Associations are transferred fractionally, such that the pair with the highest similarity receives the bulk of the score. The relation is not linear: empirically analysis (not shown) suggests that competing similarity links should be down weighted, relative to the best link, as follows: (i) express similarities as values between zero and one, i.e. normalize by self-hit; (ii) transform similarities using $s' = \exp(-k_1 s)$, thereby amplifying their ‘spread’; (iii) re-normalize so that, between the two species, all similarities for a protein family add up to one; (iv) each pair of proteins, A and B in the target species now receives a share of the association score: $S_{target} = S_{source} \cdot k_2 \cdot s'_{A} \cdot s'_{B}$. (optimal values for $k_1$ and $k_2$ were empirically found to be 0.7 for both).

Table 1. The number of associations stored in STRING, shown separately for each data source and confidence range (low confidence: scores <0.4; medium: 0.4 to 0.7; high: >0.7)

<table>
<thead>
<tr>
<th>Association evidence type</th>
<th>Low confidence</th>
<th>Medium confidence</th>
<th>High confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conserved neighborhood</td>
<td>441,385</td>
<td>111,785</td>
<td>32,401</td>
</tr>
<tr>
<td>Gene fusions</td>
<td>33,578</td>
<td>3,675</td>
<td>4,393</td>
</tr>
<tr>
<td>Phylogenetic co-occurrence</td>
<td>10,634,199</td>
<td>1,091,977</td>
<td>175,774</td>
</tr>
<tr>
<td>Co-expression</td>
<td>4,556,189</td>
<td>615,825</td>
<td>29,879</td>
</tr>
<tr>
<td>Database imports</td>
<td>38,683</td>
<td>11,193</td>
<td>7,695</td>
</tr>
<tr>
<td>Large-scale experiments</td>
<td>207,829</td>
<td>26,226</td>
<td>3,401</td>
</tr>
<tr>
<td>Literature co-occurrence</td>
<td>413,521</td>
<td>154,594</td>
<td>55,387</td>
</tr>
<tr>
<td>Combined score</td>
<td>15,498,524</td>
<td>1,937,091</td>
<td>368,669</td>
</tr>
</tbody>
</table>

The confidence increases when methods are combined (e.g. there are more high-confidence links in the last row than the simple sum). Data from version 5.1 of STRING.

ACKNOWLEDGEMENTS

This work was supported in part by grants from the Bundesministerium für Forschung und Bildung, Germany, from the Netherlands Organization of Scientific Research (NOW), and from The Knut and Alice Wallenberg Foundation (to S.D.H.).

REFERENCES


