nature machine intelligence

ARTICLES

https://doi.org/10.1038/s42256-021-00419-7

Check for updates

Protein function prediction for newly sequenced organisms

Mateo Torres[1,4], Haixuan Yang[2,4], Alfonso E. Romero[3,4] and Alberto Paccanaro[1,3]✉

Recent successes in protein function prediction have shown the superiority of approaches that integrate multiple types of experimental evidence over methods that rely solely on homology. However, newly sequenced organisms continue to represent a difficult challenge, because only their protein sequences are available and they lack data derived from large-scale experiments. Here we introduce S2F (Sequence to Function), a network propagation approach for the functional annotation of newly sequenced organisms. Our main idea is to systematically transfer functionally relevant data from model organisms to newly sequenced ones, thus allowing us to use a label propagation approach. S2F introduces a novel label diffusion algorithm that can account for the presence of overlapping communities of proteins with related functions. As most newly sequenced organisms are bacteria, we tested our approach in the context of bacterial genomes. Our extensive evaluation shows a great improvement over existing sequence-based methods, as well as four state-of-the-art general-purpose protein function prediction methods. Our work demonstrates that employing a diffusion process over networks of transferred functional data is an effective way to improve predictions over simple homology. S2F is applicable to any type of newly sequenced organism as well as to those for which experimental evidence is available. A free, easy to run version of S2F is available at https://www.paccanarolab.org/s2f.

Fewer than 1% of the available protein sequences are currently annotated with reliable information, and the gap between unannotated and annotated sequences is widening at an unprecedented rate[1] (Supplementary Note 1). Traditional experimental approaches to determine protein function are usually expensive, time-consuming and provide low throughput. Although higher-throughput approaches have recently been developed, they are also proving to be insufficient to cope with the sheer number of new sequences produced by next-generation sequencing techniques[2]. In this context, the computational annotation of protein function has become a crucial step in achieving a better understanding of the complex mechanisms of living cells.

Newly sequenced organisms represent a particularly difficult challenge for automated annotation methods because only their protein sequences are available and, in general, we lack any other data derived from large-scale functional experiments. In fact, protein function prediction is somewhat easier for more studied organisms, including model organisms, where multiple types of functional experimental evidence (for example, gene expression, proteomics data) are available that can be integrated with sequence information. The Critical Assessment of Functional Annotation Challenge (CAFA)[3] has indeed shown that advanced methods that integrate multiple types of information for the prediction of Gene Ontology (GO)[4] terms substantially outperform methods that use only sequence information.

Network propagation approaches have been shown to be among the most successful methods to predict protein function when some sort of experimental evidence is available[5]. These methods combine and amplify existing knowledge about the function of some of the proteins by propagating it through networks where nodes represent proteins and edges represent pairwise functional relationships between them that are derived from experiments (for example, physical interaction, co-occurrence in protein complexes, co-expression). In other words, these methods expand an initial set of functional labels available for some experimentally characterized proteins (seeds) to related neighbouring proteins, thus exploiting the guilt-by-association principle, according to which highly connected nodes should share similar functional properties. However, until now, these ideas could not be applied to newly sequenced organisms, because in this case both the seeds and the networks are unavailable.

This Article introduces S2F (Sequence to Function), a novel network propagation-based method for the functional annotation of newly sequenced organisms. Our main idea is to systematically transfer functionally relevant data that are available for model organisms to newly sequenced organisms, thus allowing us to use network propagation to predict protein function. S2F presents a novel network propagation algorithm that can account for the presence of overlapping communities of proteins with related functions.

Because most newly sequenced organisms are bacteria, we have developed and tested our solutions in the context of bacterial genomes. The Bacteria superkingdom is also the one with the most available sequenced proteins in UniProtKB (Supplementary Note 1), and the functional characterization of bacteria holds great potential in fields ranging from alternative energy sources to understanding and treating disease. However, the ideas presented here are more widely applicable to protein function prediction for any type of organism, and an earlier version of our algorithm has successfully been applied to organisms from other kingdoms[3,6].

Results

The aim of S2F is to predict the function of each of the proteins in a newly sequenced organism. Functional categories are defined according to the GO[4], where terms are organized in a hierarchical structure with several domains and levels of specificity. The prediction of protein function is a multi-class, multi-label classification problem—multi-class, as there are over 40,000 possible GO terms

[1]Escola de Matemática Aplicada, Fundação Getulio Vargas, Rio de Janeiro, Brazil. [2]School of Mathematics and Statistical Sciences, National University of Ireland Galway, Galway, Ireland. [3]Department of Computer Science, Centre for Systems and Synthetic Biology, Royal Holloway, University of London, Egham, UK. [4]These authors contributed equally: Mateo Torres, Haixuan Yang, Alfonso E. Romero. ✉e-mail: alberto.paccanaro@rhul.ac.uk

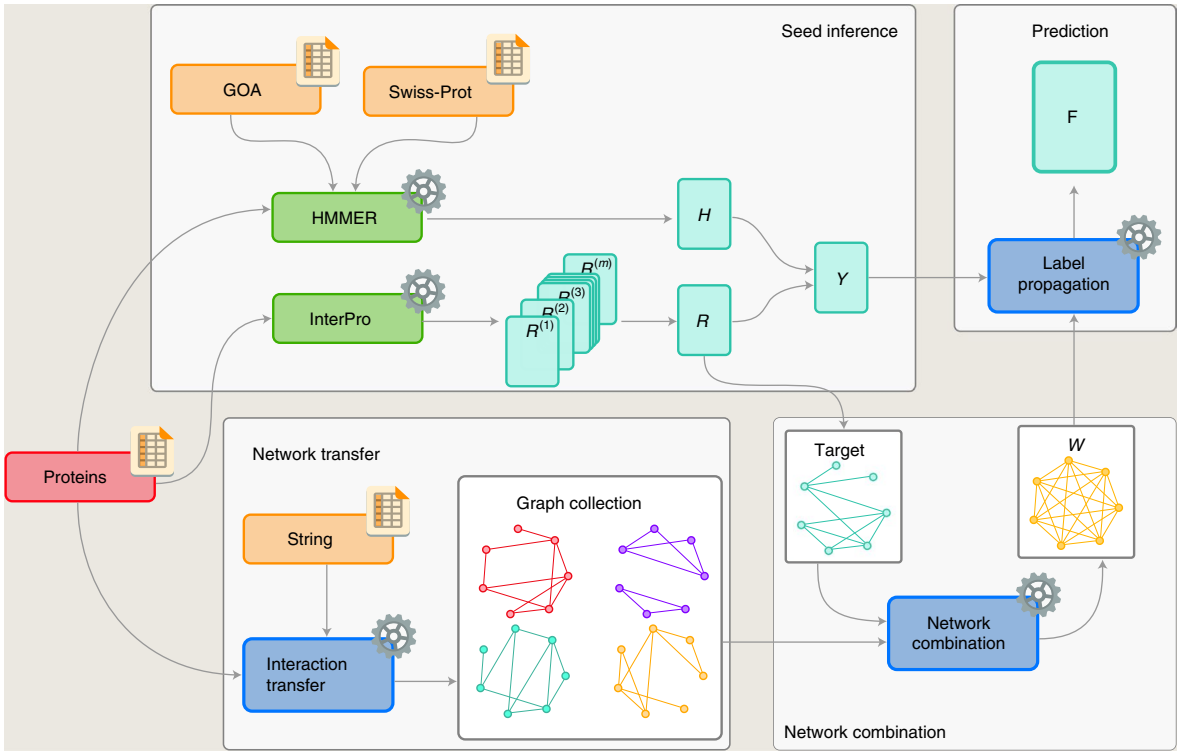

Fig. 1 | Overview of the S2F approach. The set of n protein sequences of the target organism (shown in red) constitutes the input to the system, and t is the total number of GO terms to be predicted. External datasets (STRING[15], GOA[17] and UniProtKB[27]) are shown in orange. Seed inference: on running HMMER on the input sequences against experimentally annotated sequences from UniProtKB/Swiss-Prot we obtain an $(n \times t)$ matrix H of predictions (the HMMER seed set). Running InterPro we obtain m matrices of predictions $R^{(m)}$, one per InterPro model, each of size $(n \times t)$. These matrices are then combined into a single $(n \times t)$ matrix R (the InterPro seed set). The combined seed set Y, which will be used for the label propagation, is a linear combination of H and R. Network transfer: a collection of networks is built by our interaction transfer procedure using known functional relationships between proteins in every organism from the STRING database. Network combination: transferred networks are linearly combined into a single network W. The weights of the linear combination are learned using an auxiliary target network built from R. Prediction: the network W and seed set Y are fed into our label propagation algorithm, which outputs the protein function prediction F, an $(n \times t)$ matrix where each row corresponds to a protein, each column corresponds to a GO term and each entry F_{ij} is related to the probability for protein i to have function j. For a given protein i, its labels F_i are guaranteed to be consistent; that is, they satisfy the GO 'true path rule'.

that can be annotated to a protein, and multi-label, because each protein can be annotated with multiple GO terms. Importantly, the hierarchical structure of the GO must be taken into account for the prediction, because whenever a protein is annotated with a GO term, it is also annotated with all its ancestor terms up to the root of the ontology (this is known as the 'true path rule'[7,8]). Therefore, an important requirement for the output of any protein function prediction method is to be consistent: if a GO term is predicted with a certain probability, its parent terms must be predicted with an equal or greater probability[9].

S2F consists of four main components (Fig. 1):

(1) A method to infer the initial seeds, which combines the output of InterPro[10] and HMMER (http://hmmer.org/, version 3.1b2) to obtain a set of initial predictions that is consistent
(2) A method for network transfer, which relies on the concept of interolog[11,12] to infer several functional networks
(3) A method for network combination, which combines the different functional networks into a single one
(4) A label propagation algorithm, which diffuses the seed information to obtain a prediction

In the following, we will describe each component in turn. We will assume that we wish to predict the function for a newly sequenced organism (target organism) with n proteins, and that the GO contains t terms.

S2F seed inference. InterPro[10] constitutes an excellent starting point for predicting protein function from its sequence as it provides predictions from 14 different protein signature databases. We consolidate its output into an $n \times t$ matrix of predictions R (Methods) that is consistent, and where each entry R_{ij} is the fraction of InterPro models in which the (i, j) association is present.

Although InterPro predictions are extremely accurate, they are often limited in number and involve only a few GO terms. To enrich the catalogue of GO terms that appear in our initial seed set, HMMER (http://hmmer.org/, version 3.1b2) is run for every protein in the target organism against the experimentally annotated sequences in UniProtKB/Swiss-Prot (Supplementary Note 2). This results in the HMMER seed set, a binary matrix H of size $(n \times t)$, which is then up-propagated according to the true path rule[7,8]. A convex combination of H and R gives us the consistent combined seed set $Y \in \mathbb{R}^{n \times t}$:

$$Y = \alpha R + (1 - \alpha) H$$

where $\alpha \in \mathbb{R}$, $0 \leq \alpha \leq 1$ controls the relative contribution of InterPro and HMMER predictions, and each entry of Y, $0 \leq Y_{ij} \leq 1$.

S2F network transfer. We build networks where nodes represent target organism proteins and edges represent pairwise functional relationships (interactions) between them. Because experimental

Table 1 | List of bacteria that satisfy the selection criteria, with number of genes and annotations

NCBI ID	Name	Genes	Experimentally annotated genes	BP terms with >3 annotations	MF terms with >3 annotations	CC terms with >3 annotations
272624	*Legionella pneumophila* subsp. *pneumophila Philadelphia 1*	2,076	18	30	8	8
223283	*Pseudomonas syringae* pv. *tomato*	5,055	25	48	32	15
359391	*Brucella abortus*	2,229	26	17	8	14
99287	*Salmonella typhimurium*	3,764	116	183	46	24
198628	*Dickeya dadantii*	3,411	102	214	21	13
1111708	*Synechocystis* sp.	2,442	137	101	21	30
224308	*Bacillus subtilis*	3,410	375	301	120	24
208964	*Pseudomonas aeruginosa*	4,487	947	695	222	42
83332	*Mycobacterium tuberculosis*	3,284	1,027	797	280	45
83333	*Escherichia coli*	3,906	3,350	1,546	706	134

The number of terms with more than three annotations in each of the GO domains is calculated after up-propagation and therefore may be larger than the number of experimentally annotated genes. NCBI, National Center for Biotechnology Information; BP, biological process; MF, molecular function; CC, cellular component.

evidence of functional relationships between proteins is not available for newly sequenced organisms, to create these networks we exploit the fact that these relationships are often conserved across species[13,14]. This allows us to transfer existing evidence from well-studied organisms to newly sequenced ones.

Our starting point is the seminal work by Yu et al.[12], who transferred different types of functional network with high precision using the concept of interolog-mapping first proposed by Walhout and others[11]. The idea is that, given two proteins A and B in the target organism, if there exists a pair of proteins A′ and B′ that are known to interact in another organism (source organism), such that A is an orthologue of A′ and B is an orthologue of B′, then we can infer an interaction between A and B.

Our transfer algorithm derives from the one proposed by Yu et al.[12] (details are provided in Methods and Supplementary Note 3). S2F uses STRING[15] as the dataset of different types of experimental interactions in source organisms. For each type of interaction, S2F builds one transferred network, r, that can be represented as a matrix $W^{(r)} \in \mathbb{R}^{n\times n}$, where each entry $W_{ij}^{(r)}$ represents the strength of the interaction between proteins i and j in r. For a given target organism, S2F transfers five types of interaction, namely 'neighborhood', 'experiments', 'co-expression', 'textmining' and 'database', using the experimental interactions available for any organism in STRING.

S2F network combination. Having obtained a set of transferred networks for the target organism, we now face the task of combining them into a single network for diffusing the seeds. Our approach is to linearly combine the different networks through learned coefficients. These coefficients provide us with interesting information about the relative importance and role of each network in the prediction. Although other systems learn this combination (for example, GeneMANIA[16]), the solution we propose here is applicable to our problem, where no initial set of known labels is available.

We begin by using the InterPro predictions to build a network of functional similarities $T \in \mathbb{R}^{n\times n}$, where the similarity between proteins i and j, T_{ij}, is defined as

$$T_{ij} = \frac{|N_i \cap N_j|}{|N_i \cup N_j|}$$

where N_i and N_j are the sets of all GO terms above a threshold τ that are associated to proteins i and j, respectively, in R; that is, $N_i = \{k|R_{ik} > \tau\}$ and $N_j = \{k|R_{jk} > \tau\}$. Therefore, T_{ij} is the Jaccard similarity between sets of GO terms that are assigned by InterPro to proteins i and j.

Given p networks $W^{(r)}$ with $r \in \{1, \ldots, p\}$, we combine them into a single network $W \in \mathbb{R}^{n\times n}$ using a weighted linear combination, where the vector of weights $\hat{\mathbf{c}} \in \mathbb{R}^p$ is learned by minimizing the square of the difference between T and the linear combination (Methods).

S2F label propagation. Proteins rarely perform their functions in isolation, but rather they act as part of functional groups. As mentioned earlier, network propagation methods for protein function prediction exploit exactly this fact—groups of proteins that are highly connected in functional networks form communities that share a similar function. Importantly, when a protein has more than one function, it will belong to more than one such functional group. We notice that such proteins, lying at the intersection of communities, are, in general, more functionally similar compared to their neighbours, because they share more functional roles. Therefore, when a set of proteins has more than one function, the propagation of information (or diffusion) between proteins within this set should be higher than the diffusion between proteins in this set and proteins outside this set. However, this does not happen with existing diffusion methods (for details see Supplementary Note 6). Here we propose a novel label propagation method that explicitly models overlapping communities and, in this way, corrects this problem.

We begin by defining the matrix $W^{\mathrm{S2F}} \in \mathbb{R}^{n\times n}$, a transformation of the combined network W whose entry W_{ij}^{S2F} is defined as

$$W_{ij}^{\mathrm{S2F}} = \frac{1}{2}\left(\frac{1}{d_i} + \frac{1}{d_j}\right) J_{ij} W_{ij}$$

where $d_i = \sum_{j=1}^{n} J_{ij} W_{ij}$ and J is a weighted Jaccard similarity matrix that models the overlapping community effect (Methods). We also define a diagonal matrix D^{S2F} where the ith diagonal element $D_{ii}^{\mathrm{S2F}} = \sum_j W_{ij}^{\mathrm{S2F}}$. Our algorithm produces a prediction matrix $F \in \mathbb{R}^{n\times t}$ for all the n proteins of the organism and all the t GO terms by computing the following:

$$F = (I + \lambda L)^{-1} Y$$

where Y is the matrix containing the initial labelling, I is the identity matrix, $L = D^{\mathrm{S2F}} - W^{\mathrm{S2F}}$ is the Laplacian of W^{S2F}, and $\lambda > 0$ is the regularization parameter (Methods).

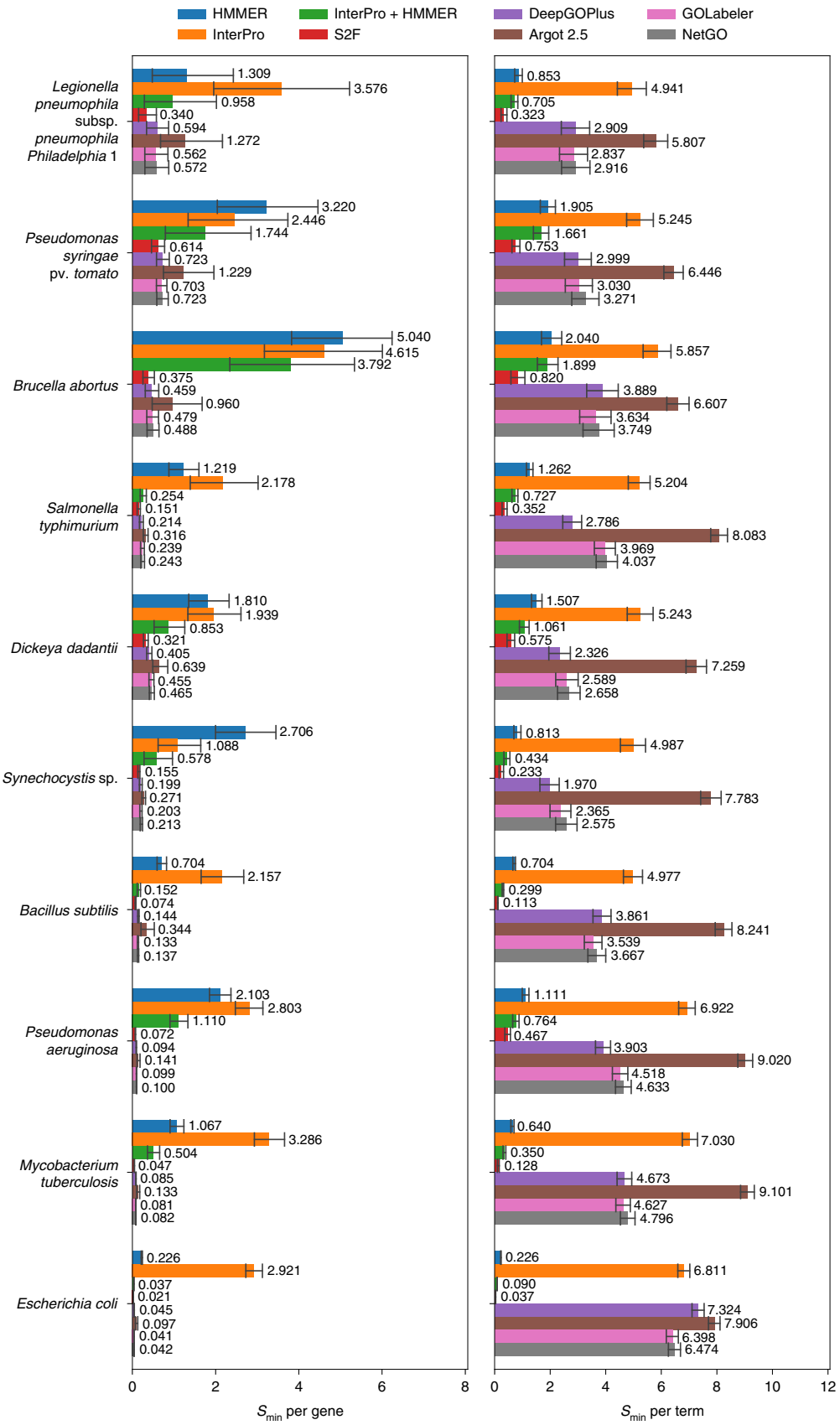

Fig. 2 | S_{min} metric for every organism per gene and per term, with lower values being better. Comparison of HMMER, InterPro, HMMER + InterPro, S2F, Argot 2.5, DeepGOPlus, GOLabeler and NetGO. Values indicate the mean of the metric over genes (left) or terms (right), and error bars indicate a confidence interval of 95%, estimated using 10,000 bootstrap iterations on the gene set or term set, respectively.

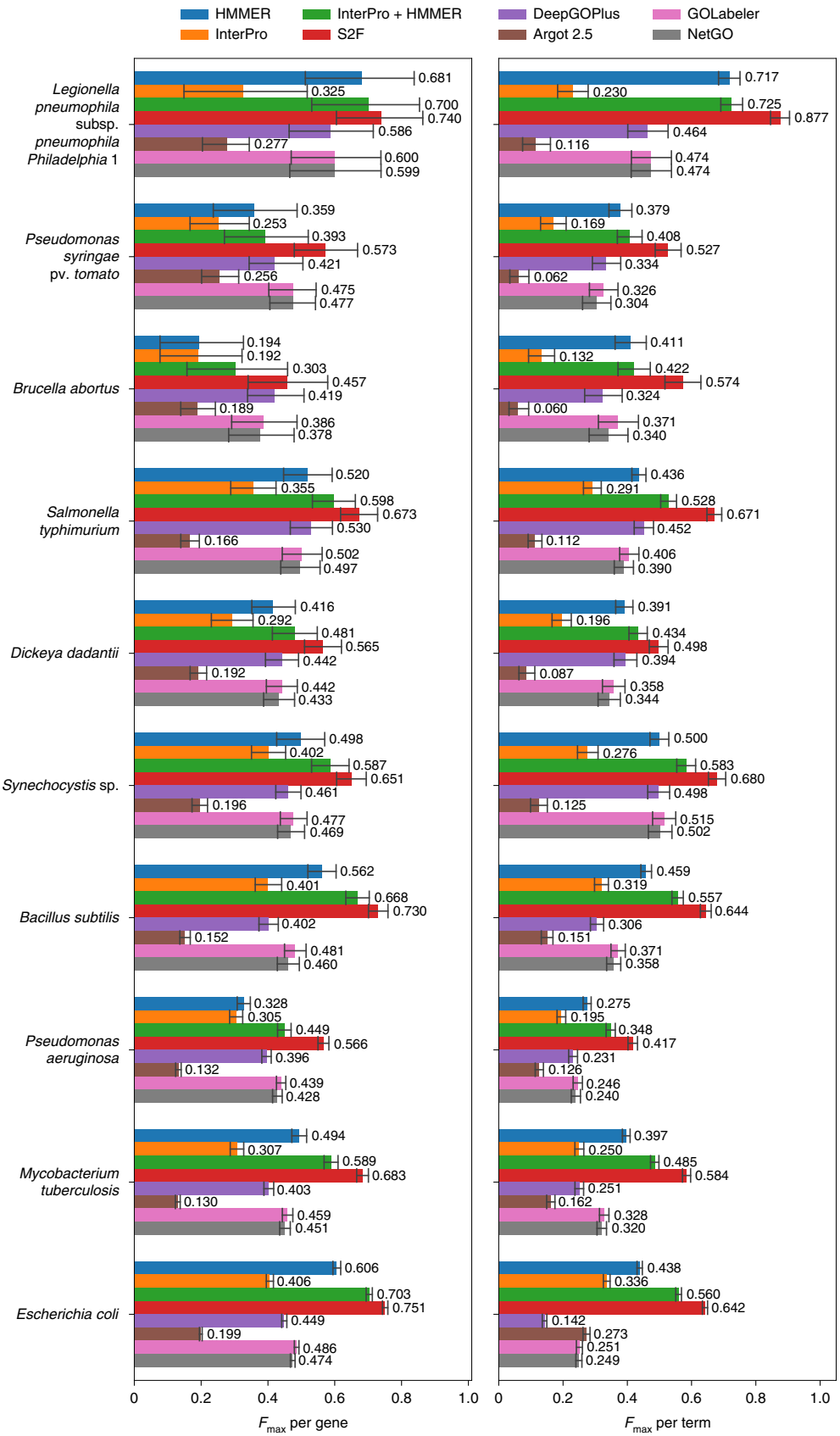

Fig. 3 | F_{max} for every organism per gene and per term, with higher values being better. Comparison of HMMER, InterPro, HMMER + InterPro, S2F, Argot 2.5, DeepGOPlus, GOLabeler and NetGO. Values indicate the mean of the metric over genes (left) or terms (right), and error bars indicate a confidence interval of 95%, estimated using 10,000 bootstrap iterations on the gene set or term set, respectively.

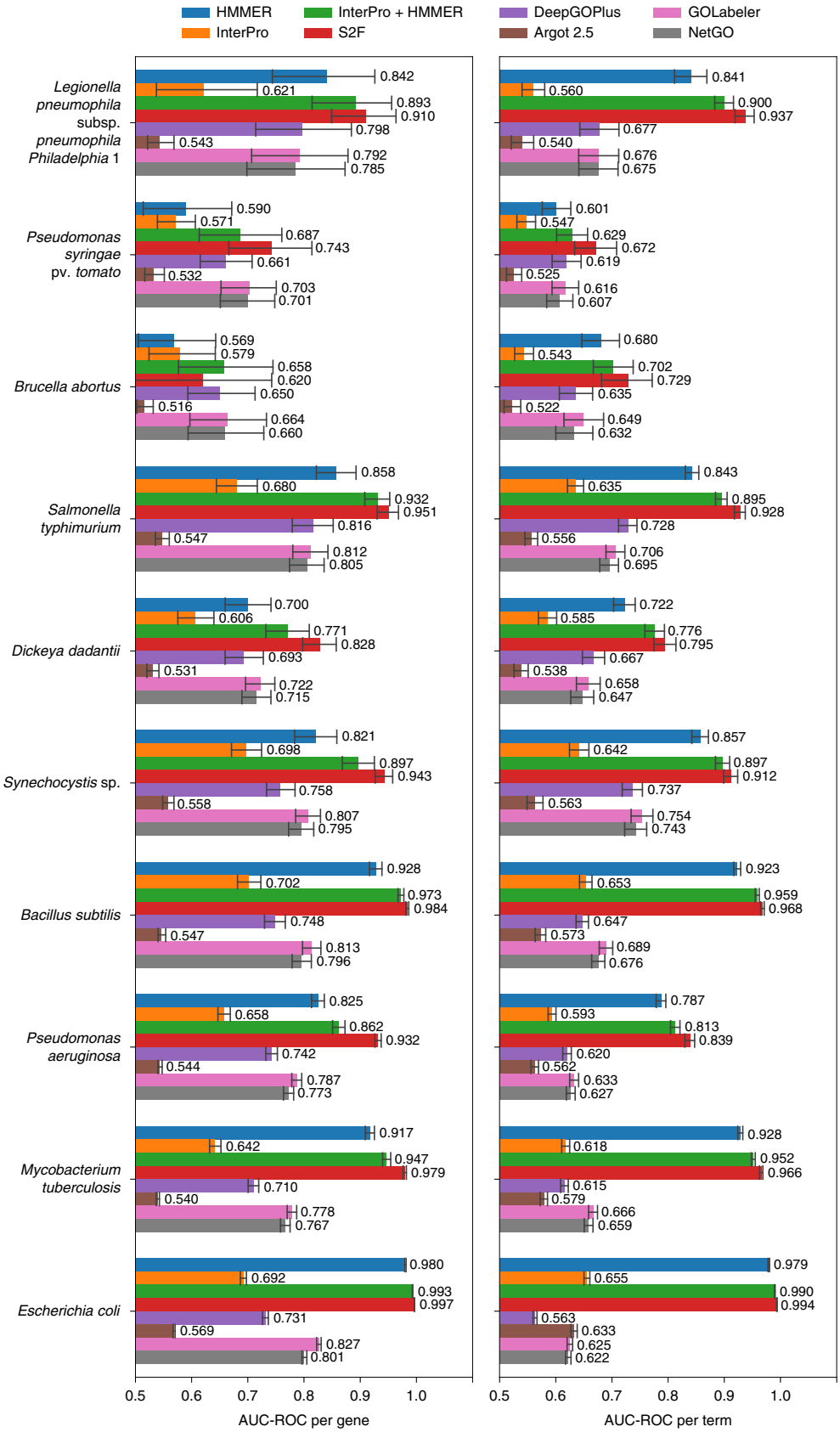

Fig. 4 | AUC-ROC for every organism per gene and per term, with higher values being better. Comparison of HMMER, InterPro, HMMER + InterPro, S2F, Argot 2.5, DeepGOPlus, GOLabeler and NetGO. Values indicate the mean of the metric over genes (left) or terms (right), and error bars indicate a confidence interval of 95%, estimated using 10,000 bootstrap iterations on the gene set or term set, respectively.

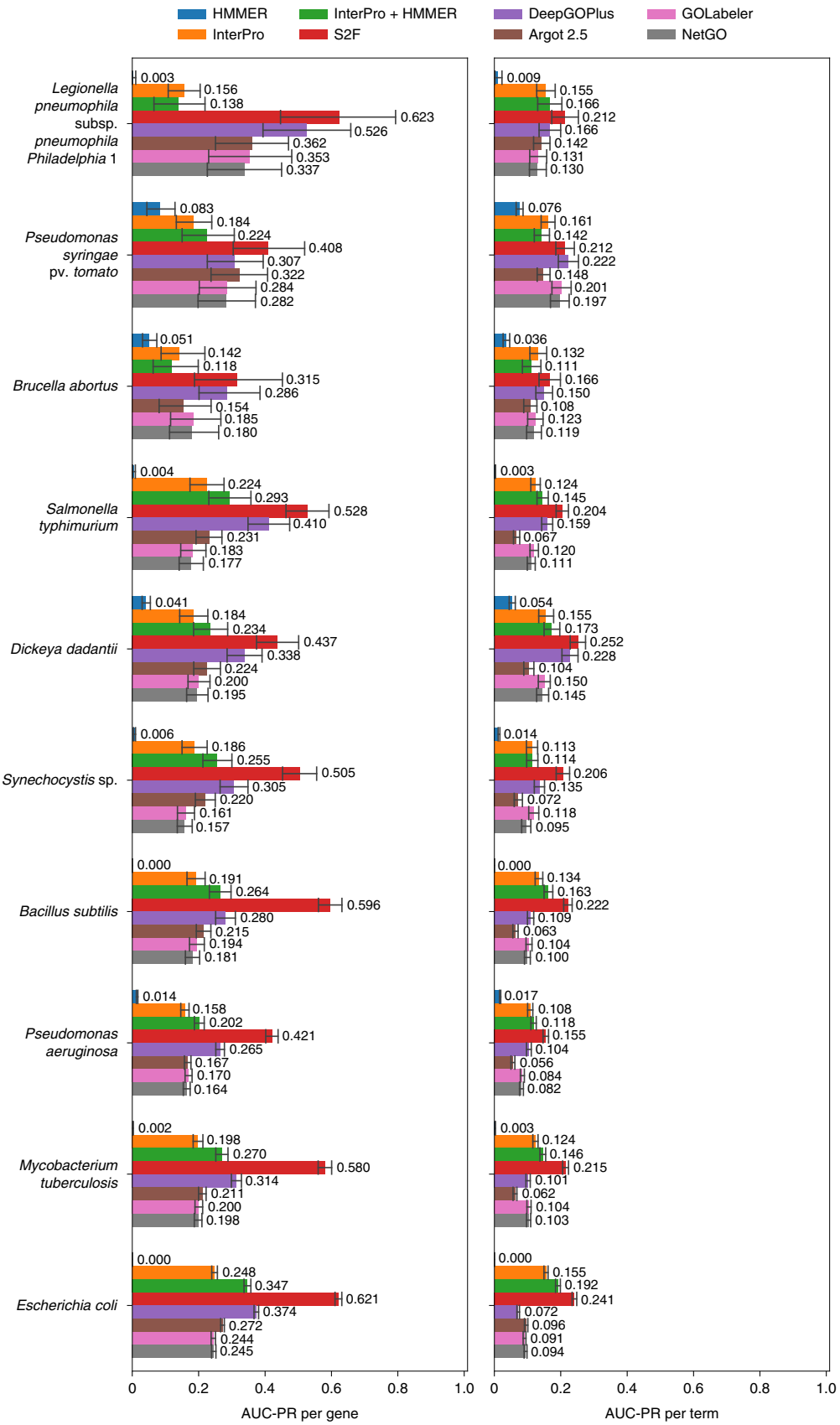

Fig. 5 | AUC-PR for every organism per gene and per term, with higher values being better. Comparison of HMMER, InterPro, HMMER + InterPro, S2F, Argot 2.5, DeepGOPlus, GOLabeler and NetGO. Values indicate the mean of the metric over genes (left) or terms (right), and error bars indicate a confidence interval of 95%, estimated using 10,000 bootstrap iterations on the gene set or term set, respectively.

We show that this label propagation algorithm does not suffer from the problem described above for overlapping communities (Supplementary Note 6). Moreover, we prove that it satisfies the necessary conditions to ensure that, for each pair of terms j and k such that j is an ancestor of k (in these cases $Y_{ij} \geq Y_{ik}$ for every i), we have that $F_{ij} \geq F_{ik}$ for every i (the proof is provided in Supplementary Note 7). As a consequence, because Y is consistent with the GO structure, F will also be consistent.

Experimental set-up

We present the evaluation of S2F on bacteria from UniProtKB. Following the evaluation procedure used by most authors[3,6] the performance of S2F in predicting protein function was assessed both in a per-gene and in a per-term setting. In per-gene predictions, given a gene, we assess the performance of S2F at predicting a set of functions associated to that gene. Conversely, in per-term predictions, given a function, we assess the performance of S2F at predicting a set of genes that perform that function.

The performance was assessed against a set of known experimental annotations. Therefore, the bacteria used for testing were chosen so that they had at least a few experimentally annotated genes (to be able to assess the performance in a per-gene setting) while maintaining a reasonable diversity of annotated GO terms (to be able to assess the performance in a per-term setting) in the GOA database[17]. The ten bacteria in Table 1 satisfied our set of criteria (the criteria are detailed in Methods).

This set of bacteria provides a good testbed for our experiments. The amount of experimental annotations in these bacteria covers a wide spectrum, ranging from well-studied bacteria (for example, *Escherichia coli*) to more obscure ones that are not even included in STRING (for example, *Brucella abortus*).

In our experiments, we tested the performance at predicting the functional annotation for the whole genome for each of the ten bacteria, in turn. To avoid circular reasonings, when testing each bacterium, we carefully removed any functional information for that bacterium as well as for any phylogenetically close species. To do this, for each bacterium, we created a list of excluded species in two steps. First, starting from that bacterium, we navigated the NCBI taxonomy moving up two levels (that is, to the parent of the parent node) and we included in our list that node and all its descendants. Second, we added to the list all the nodes in the NCBI taxonomy that had a similar name. Having created a list of excluded species, we removed any information about these species from STRING, as well as information about their proteins from the GOA database. The detailed list of all organisms excluded when testing each specific bacterium is provided in the Supplementary Data.

Predicted annotations were evaluated against the existing functional annotations (GOA files in Supplementary Data) using the well-established metrics that have been used in the CAFA challenge[3]: The maximum F measure F_{max}, the minimum semantic distance S_{min}, and the areas under the receiver operating characteristic (ROC) and precision-recall (PR) curves AUC-ROC and AUC-PR metrics (for details see Supplementary Note 12).

Evaluation

We compared the performance of S2F against InterPro, HMMER, Argot 2.5[18], DeepGOPlus[19], GOLabeler[20] and NetGO[21]. InterPro and HMMER are among the best and most widely used sequence-based methods for predicting protein function for newly sequenced organisms. The other four methods, although not explicitly conceived for this problem, could nevertheless be employed here as they are able to predict protein function using sequence information alone. Argot 2.5[18] and GOLabeler[20] were among the top performers in the last edition of the CAFA competition[6]. NetGO[21] and DeepGOPlus[19] were introduced after the last CAFA competition and were shown to perform very well against the top CAFA algorithms. (Details of the implementation, parameter settings and a description of these algorithms are provided in the Methods and Supplementary Notes 14, 16 and 17).

Figures 2–5 show the AUC-ROC, AUC-PR, F_{max} and S_{min} evaluated per gene and per term for S2F and each competitor algorithm. (An interactive version of these results is also available in the result explorer on our website, https://www.paccanarolab.org/s2f). We can see that S2F outperformed the other methods according to the vast majority of performance measures for the ten bacteria—it is surpassed only in 4 of the 80 bacteria–measure combinations, most often on the AUC-ROC measure. To better appreciate the increase in performance offered by S2F, we also explicitly report the percentage of improvement of S2F versus each competitor for each of the ten organisms (Supplementary Figs. 53–59 in Supplementary Note 15).

Analysing these results, we see that, as expected, the accuracy of the S2F predictions does depend on the accuracy of InterPro and HMMER, which provide the initial seeds for the S2F diffusion process. An interesting question is whether the improved performance of S2F is merely due to the fact that it combines the labels of InterPro and HMMER, or whether the diffusion of these labels through the transferred networks has a role in its performance. For this reason, we also report in the figures the performance of the linear combination of InterPro and HMMER labels that we used as seeds for the diffusion process in S2F (matrix Y). We can see that, with the exception only of the AUC-ROC for *Brucella abortus*, S2F shows an improvement when compared with the simple linear combination of the InterPro and HMMER outputs. This means that S2F is able to effectively combine the information of these labels together with the evolutionary information contained in the interolog graphs.

As we mentioned earlier, by integrating InterPro and HMMER we aimed to obtain seeds that combined the high accuracy and specificity offered by InterPro with the high coverage provided by HMMER. To check whether our linear combination, controlled by the parameter α, achieved this, we analysed how the different setting of α affected the S2F results (details of the experiments are described in Supplementary Note 13). Supplementary Figs. 48–51 show that, in general, a combination of InterPro and HMMER seeds ($0 < \alpha < 1$ gives much better results in terms of S2F performance than when using only seeds from either of them ($\alpha = 0$ or $\alpha = 1$). However, just looking at S2F performance, it is unclear how to set the value of α, as there is disagreement among the different performance measures and organisms. At the same time, an important objective in real-world scenarios is to predict, for a given gene, a small set of terms that are highly accurate while being as specific as possible. Therefore, we analysed the information content of the top genes predicted by S2F for different values of α (Supplementary Fig. 52). Our results show that, in this scenario, high values of α (for example, $\alpha = 0.9$) should be preferred.

We also evaluated the predictions obtained by diffusing the outputs of InterPro and HMMER, separately, on the interolog network W. Supplementary Figs. 11–14 (Supplementary Note 8; also available in the interactive data explorer on our website, https://www.paccanarolab.org/s2f) show how our diffusion process is able to improve the labels obtained by InterPro (or HMMER). This means that our diffusion on combined interolog networks is an effective way to improve protein function prediction over simpler homology methods.

Our diffusion method was motivated by our desire to model the presence of overlapping communities in functional networks. It is unclear how to quantify exactly the number of proteins being shared across communities, as this is obscured by the relationships among functional labels as well as the noise and incompleteness of available annotations. However, the semantic similarity of proteins with known function can provide some insight, as we can quantify the correlation between the graph onto which we diffuse, W^{S2F}, and a graph of semantic similarities among functionally annotated

proteins, G^{SS}. Supplementary Fig. 17 shows the values of these correlations for each of the ten bacteria and compares them with correlations between G^{SS} and W^{GM}, the graph used by GeneMANIA[16], a diffusion-based method for protein function prediction in model organisms that does not explicitly model overlapping communities (for details of these experiments see Supplementary Note 6). We can see that W^{S2F} shows higher correlation with the semantic similarity graph G^{SS} in the great majority of the cases, for different organisms and across different GO ontologies.

Finally, to further demonstrate how S2F can facilitate biological research by generating feasible hypotheses, we performed a prospective evaluation. We deployed S2F to make predictions using only data available up to December 2014 and we assessed its accuracy on proteins that were experimentally annotated between 2015 and 2021. The experiments are detailed in Supplementary Note 11. Supplementary Figs. 44–47 show that, although the performance of InterPro is relatively stable, for some bacteria the overall performance of HMMER (and, as a consequence, of the InterPro+HMMER combination) seems to worsen greatly. As expected, the performance of S2F decreases in these cases, but overall the diffusion process is able to alleviate the effect and compensate for the lower quality of the seeds.

Discussion

The difficulty of protein function prediction, one of the most important problems in computational biology, varies greatly depending on how much experimental information is available for the organism under investigation. Predictions for well-studied organisms can rely on multiple types of functional experimental evidence (for example, gene expression and proteomics data) that can be represented in the form of graphs. For these organisms, network propagation approaches that amplify existing knowledge about the function of some of the proteins have been shown to be very effective[5,16,22,23].

This Article introduces S2F, a method that applies a network propagation algorithm to organisms for which only sequence information is available. The main idea is to create networks of interologs by systematically transferring functional data that are available for model organisms and to use these networks to combine and amplify a few preliminary GO labels (seeds) obtained through homology or identifiable protein features.

Our work shows that employing a diffusion process over networks of interologs is an effective way to improve predictions over simple homology. The improvement comes from combining information: S2F effectively integrates homology information and identifiable protein features (preliminary GO labels from HMMER and InterPro) together with evolutionary information contained in the interolog graphs, through a diffusion process. S2F includes a novel network propagation algorithm that can account for the presence of overlapping communities of nodes with related functions.

Ultimately, the accuracy of S2F when predicting the function for a specific organism will depend on several factors, including the specificity and diversity of the preliminary GO labels, as well as the density of the interolog networks, which in turn depends on the evolutionary distance from organisms with existing functional experimental evidence. When predicting a GO term for a specific gene, these factors affect how many neighbours that gene has, how many of these genes have preliminary GO labels and how accurate these labels are. These factors are highly interleaved, and it is difficult to quantify the effect of each one individually. For example, it would seem reasonable to expect that S2F would generate better predictions for more highly connected nodes. We tested this hypothesis by measuring the correlation between node degrees and the performance measures for the bacteria in this study. However, our results show that the correlation was either weak and negative or not statistically significant (Supplementary Note 10 and Supplementary Figs. 26–36).

The different interolog networks that we combine are extremely sparse, with virtually no overlap among them (Supplementary Figs. 3 and 4). In this scenario, in terms of prediction performance, different combination methods would give results that are as good as the simple average of the networks (Supplementary Figs. 5–8 compare our combination strategy, the network combination used by STRING[15] and the simple average). However, our approach allows the linear combination of the different networks through learned coefficients, providing us with information about the relative importance and role of each network in the prediction (Supplementary Note 4). Our combination method is similar to the one used in GeneMANIA, but it allows us to learn these linear weights without relying on an initial set of known functional labels.

We note that the removal of functional information regarding each bacterium and its phylogenetically close species makes this problem much harder than the one tested in the regular CAFA competition settings. For this reason, the performances for Argot 2.5[18], DeepGOPlus[19], GOLabeler[20] and NetGO[21] seem generally lower than those reported earlier. Also, methods that are able to integrate global and local information seem to perform better than local methods in our setting. This can be seen by comparing the results obtained by S2F and the 'consistency method' (CM)[24]—another method that integrates global information—with the results obtained by NetGO, where the use of network information is limited locally to nodes that are just one link away from the query node. A performance comparison between our label propagation method and the CM is available in Supplementary Note 9.

In this Article we have focused and presented results for bacteria, but S2F can be applied to any organism, independently of how well functionally characterized it is. An earlier version of S2F optimized to use existing functional evidence for target organisms was submitted to the CAFA2 challenge[3], where it ranked as one of the top performing methods.

The code for S2F is freely available at https://www.paccanarolab.org/s2f. The S2F software is fast, robust and easy to set up and run. The software is fully documented, including a wiki with instructions for common use cases, instructions on how to use S2F to predict function for newly sequenced bacteria and details on how to replicate all our results, together with the necessary input data (Supplementary Data).

Methods

S2F seed inference. InterPro produces m binary matrices of predictions $R^{(k)}$, each of size $(n \times t)$ (here $k \in \{1, \ldots, m\}$ and $m \leq 14$ is the number of models for which InterPro gives at least one prediction for the target organism). To combine these matrices while ensuring that the combination is consistent with the hierarchical structure of GO, we first up-propagate these associations according to the true path rule[7,8], considering both the 'is_a' and 'part_of' relations. Each matrix $R^{(k)}$ is up-propagated separately, so any convex combination of the up-propagated matrices will be consistent. We combine them to obtain a consistent InterPro seed set $R \in \mathbb{R}^{n \times t}$ where each entry of R, R_{ij}, is defined as

$$R_{ij} = \frac{\sum_{k=1}^{m} R_{ij}^{k}}{m}$$

S2F network transfer. STRING[15] is a database that compiles several 3,123,056,667 interactions between proteins in 5,090 organisms. Interactions are divided into seven types: 'neighborhood', 'fusion', 'co-occurrence', 'experiments', 'co-expression', 'textmining' and 'database'. Each interaction is annotated with a score that ranges from 0 to 1, representing the confidence that STRING assigns for the two proteins to be functionally related.

In our transfer procedure, two proteins A and A′ are considered to be orthologues if three conditions are met:

(1) They are BLAST mutual best hits, with both e-values smaller than 1×10^{-6}.
(2) The percent identity is greater than 80%—this is to avoid transference between multi-domain proteins with different domain architecture.
(3) Their 'joint identity' (geometric mean of the two percent identities) is above 60%—Yu et al.[12] showed that this condition achieves almost perfect accuracy at identifying interacting orthologues.

When the same interaction can be transferred from multiple organisms, only the one with the highest 'joint identity' is kept. The pseudocode of the algorithm for building a collection of transferred networks for the target organism is provided in Supplementary Algorithm 1 (Supplementary Note 3). S2F only considers networks with at least three edges; that is, for every interaction type in STRING, we consider the transferred network r only if $W^{(r)}$ contains at least three values.

Finally, a homology network is added to the collection of interolog networks to increase the combined network connectivity and facilitate the diffusion process. The homology network $W^{(h)}$ is defined as the negative log of the BLAST e-value for every pair of proteins.

S2F network combination. Given p networks $W^{(r)}$ with $r \in \{1, \ldots, p\}$, we combine them into a single network W using a weighted linear combination. The vector of weights $\hat{\mathbf{c}} \in \mathbb{R}^p$, and bias $\hat{b}$ are learned by minimizing

$$\left(\hat{\mathbf{c}}, \hat{b}\right) = \underset{\mathbf{c},b}{\operatorname{argmin}} \sum_{i,j} \left(b + \sum_{r=1}^{p} c_r W_{ij}^{(r)} - T_{ij} \right)^2$$

This linear regression can be solved efficiently, and we can interpret each learned coefficient c_r as representing how much each network r contributes to the combination. An analysis of these coefficients is provided in Supplementary Note 4.

S2F label propagation. The weighted Jaccard coefficient matrix J is defined elementwise as

$$J_{ij} = \frac{\sum_k W_{ik} W_{jk}}{\sum_k W_{ik} + \sum_k W_{jk} - \sum_k W_{ik} W_{jk}}.$$

Thus, the element J_{ij} relates to how much elements i and j belong to the same community in network W. For a given term k, we learn the kth column of matrix F, which we denote by F_k, by minimizing the cost function $\mathcal{Q}(F_k)$:

$$\mathcal{Q}(F_k) = \sum_{i=1}^{n} (F_{ik} - Y_{ik})^2 + \frac{\lambda}{2} \sum_{i=1}^{n} \frac{1}{d_i} \sum_{j=1}^{n} J_{ij} W_{ij} (F_{ik} - F_{jk})^2$$

Similar to the cost function used by the CM[24], ours is the sum of two terms. The role of the first term is to conserve the initial labels Y_{ik}—this term is minimized when the node labels F_{ik} are the same as the initial labels. The second term accounts for the consistency of the labels of adjacent nodes (reflecting the guilt-by-association principle)—this term is minimized when adjacent nodes have similar labels (that is, the difference between F_{ik} and F_{jk} becomes small). Note that the importance of the difference between F_{ik} and F_{jk} is proportional to $J_{ik}W_{ik}$, which models the community effect—the more i and j are connected through their neighbours, the greater their contribution to the cost function. Furthermore, notice that

$$\frac{1}{d_i} = \frac{1}{\sum_j J_{ij} W_{ij}}$$

is a normalization factor that gives to each protein in the network similar ability to influence its neighbours, independently of its degree.

The closed-form solution that minimizes $\mathcal{Q}(F_k)$ is

$$F_k^* = (I + \lambda L)^{-1} Y_k$$

where Y_k is the initial labelling, $L = D^{\mathrm{S2F}} - W^{\mathrm{S2F}}$ is the Laplacian of W^{S2F}, whose entry W_{ij}^{S2F} is defined as

$$W_{ij}^{\mathrm{S2F}} = \frac{1}{2} \left(\frac{1}{d_i} + \frac{1}{d_j} \right) J_{ij} W_{ij},$$

and D^{S2F} is a diagonal matrix where the ith diagonal element is $D_{ii}^{\mathrm{S2F}} = \sum_j W_{ij}^{\mathrm{S2F}}$.

Bacteria selection criteria and datasets. The criteria we used for selecting bacteria were as follows:

- The bacteria must have at least ten functional annotations with an experimental or curated GO evidence code (EXP, IDA, IPI, IMP, IGI, IEP, TAS or IC) in the GOA database[17].
- The bacteria must have at least eight terms annotated with at least three genes after up-propagation, for each GO subdomain—biological process (BP), molecular function (MF) and cellular component (CC).

In our experiments, we used STRING version 11.0. All sequences in FASTA format were downloaded from UniProtKB/Swiss-Prot using the taxonomy identifiers listed in Table 1. The GO annotations were downloaded from the GOA database[17]. All datasets were downloaded in April 2020. We used HMMER version 3.1b2, InterProScan version 5.42–78.0 and blastp from BLAST 2.6.0+.

Competitor algorithms. In all our experiments, to simulate a real case scenario for the problem of predicting function in newly sequenced organisms, for each bacterium we removed any functional information regarding that bacterium as well as any functional information about species that are phylogenetically close (the list of all organisms excluded is provided in the Supplementary Data).

GOLabeler[20] and its successor, NetGO[21], are only offered as web services, and use all the data available from their sources (namely GOA, STRING, UniProtKB, InterPro) for their prediction. Therefore, the results for NetGO and GOLabeler presented here were obtained running our own implementation of these systems that had been trained using datasets from which all the aforementioned functional information had been removed. All the parameters of the component models as well as the learning to rank ensemble were set using the default values suggested by the authors[20,21]. A detailed description of how to prepare the input data and how to use our implementation of these methods is provided in Supplementary Note 17.

Argot 2.5[18] was run on its web server (http://www.medcomp.medicina.unipd.it/Argot2-5/). For each bacterium, we first used BLAST and HMMER to obtain alignments between its proteins and a version of UniProtKB from which the sequences of excluded organisms (for that bacterium) were omitted. These alignments were then submitted to the Argot 2.5 web server.

DeepGOPlus[19] was run using the code from the latest stable version available (1.0.1). To remove the information from phylogenetically close organisms, we added some pre-processing steps to the input files and small corrections were made to the prediction script. A detailed guide on how to set up and run the pre-processing and prediction is described in Supplementary Note 16.

InterPro was run using InterProScan version 5.42–78.0, the output file was then processed to extract the predictions that included GO terms.

HMMER version 3.1b2 was run against a GO annotation file that was pre-processed to keep only the experimental or curated evidence codes (EXP, IDA, IPI, IMP, IGI, IEP, TAS or IC). The output file was post-processed to remove any alignment that came from an organism that had been excluded in the prediction.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The input sequence files[25] in FASTA format for all the organisms used in this paper are available at https://doi.org/10.5281/zenodo.5514323. The same URL also contains the detailed list of all organisms excluded when testing each specific bacterium.

Code availability

The code for S2F is freely available and maintained at https://www.paccanarolab.org/s2f. The exact version[26] used for this publication is available at https://doi.org/10.5281/zenodo.5513071.

Received: 8 December 2020; Accepted: 26 October 2021;
Published online: 09 December 2021

References

1. Cruz, L. M., Trefflich, S., Weiss, V. A. & Castro, M. A. A. Protein function prediction. *Methods Mol. Biol.* **1654**, 55–75 (2017).
2. Shehu, A., Barbará, D. & Molloy, K. in *Big Data Analytics in Genomics* (ed. Wong, K.-C.) 225–298 (Springer, 2016); https://doi.org/10.1007/978-3-319-41279-5_7
3. Jiang, Y. et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol.* **17**, 184 (2016).
4. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
5. Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.* **18**, 551–562 (2017).
6. Zhou, N. et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol.* **20**, 244 (2019).
7. Valentini, G. True path rule hierarchical ensembles for genome-wide gene function prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **8**, 832–847 (2011).
8. Friedberg, I. & Radivojac, P. in *The Gene Ontology Handbook* (eds Dessimoz, C. & Škunca, N.) 133–146 (Springer, 2017); https://doi.org/10.1007/978-1-4939-3743-1_10
9. Obozinski, G., Lanckriet, G., Grant, C., Jordan, M. I. & Noble, W. S. Consistent probabilistic outputs for protein function prediction. *Genome Biol.* **9**, S6 (2008).
10. Mitchell, A. L. et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **47**, D351–D360 (2019).
11. Walhout, A. J. et al. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287**, 116–122 (2000).
12. Yu, H. et al. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res.* **14**, 1107–1118 (2004).

13. Ben-Hur, A. & Noble, W. S. Kernel methods for predicting protein-protein interactions. *Bioinformatics* **21**, i38–i46 (2005).
14. Sharan, R. et al. Conserved patterns of protein interaction in multiple species. *Proc. Natl Acad. Sci. USA* **102**, 1974–1979 (2005).
15. Szklarczyk, D. et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
16. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C. & Morris, Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol.* **9**, S4 (2008).
17. Huntley, R. P. et al. The GOA database: gene ontology annotation updates for 2015. *Nucleic Acids Res.* **43**, D1057–D1063 (2015).
18. Lavezzo, E., Falda, M., Fontana, P., Bianco, L. & Toppo, S. Enhancing protein function prediction with taxonomic constraints—the Argot2.5 web server. *Methods* **93**, 15–23 (2016).
19. Kulmanov, M. & Hoehndorf, R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* **36**, 422–429 (2020).
20. You, R. et al. GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics* **34**, 2465–2473 (2018).
21. You, R. et al. NetGO: improving large-scale protein function prediction with massive network information. *Nucleic Acids Res.* **47**, W379–W387 (2019).
22. Makrodimitris, S., van Ham, R. C. H. J. & Reinders, M. J. T. Automatic gene function prediction in the 2020s. *Genes* **11**, 1264 (2020).
23. Cao, M. et al. Going the distance for protein function prediction: a new distance metric for protein interaction networks. *PLoS ONE* **8**, e76339 (2013).
24. Zhou, D., Bousquet, O., Lal, T. N., Weston, J. & Schölkopf, B. Learning with local and global consistency. In *Proc. 16th International Conference on Neural Information Processing Systems* (eds Thrun, S. et al.) 321–328 (MIT, 2004).
25. Torres, M., Yang, H., Romero, A. E. & Paccanaro, A. Input data for 'Protein function prediction for newly sequenced organisms'. *Zenodo* https://doi.org/10.5281/ZENODO.5514323 (2021).
26. Torres, M., Yang, H., Romero, A. E. & Paccanaro, A. Source code for 'Protein function prediction for newly sequenced organisms'. *Zenodo* https://doi.org/10.5281/ZENODO.5513071 (2021).
27. UniProt Consortium UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

Acknowledgements

The first idea for this project was conceived in discussions with T. Gianoulis, who we remember dearly for her intelligence, kindness, enthusiasm and passion for research. We also thank P. Bhat, T. Nepusz, J. Caceres, M. Frasca, G. Valentini, A. Devoto, L. Bögre, R. Sasidharan and M. Gerstein for many important and stimulating discussions. A.P. was supported by Biotechnology and Biological Sciences Research Council (https://bbsrc.ukri.org/) grants numbers BB/K004131/1, BB/F00964X/1 and BB/M025047/1, Medical Research Council (https://mrc.ukri.org) grant number MR/T001070/1, Consejo Nacional de Ciencia y Tecnología Paraguay (https://www.conacyt.gov.py/) grants numbers 14-INV-088 and PINV15–315, National Science Foundation Advances in Bio Informatics (https://www.nsf.gov/) grant number 1660648, Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro grant number E-26/201.079/2021 (260380) and Fundação Getulio Vargas.

Author contributions

A.P. conceived the study. A.P. and H.Y. devised the algorithms, developed the prototype and performed preliminary evaluations. M.T. and A.E.R. implemented and extended the algorithms and evaluation metrics, performed large-scale experiments and analysed the results. A.P., M.T. and A.E.R. wrote the manuscript and evaluated the biological relevance of the results. All authors discussed the results and implications. A.P. supervised the project.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-021-00419-7.

Correspondence and requests for materials should be addressed to Alberto Paccanaro.

Peer review information *Nature Machine Intelligence* thanks Jiecong Lin and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature portfolio

Corresponding author(s): Alberto Paccanaro

Last updated by author(s): Sep 17, 2021

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	In our experiments, we used STRING version 11.0 (http://string-db.org/). All sequences in FASTA format were downloaded from UniProtKB/Swiss-Prot (https://www.uniprot.org/) using the taxonomy identifiers listed for the selected bacteria. The GO annotations were downloaded from the GOA database (https://www.ebi.ac.uk/GOA/). All datasets were downloaded in April 2020. We used HMMER version 3.1b2 (http://hmmer.org/), InterProScan version 5.42-78.0 (https://www.ebi.ac.uk/interpro/), and blastp from BLAST 2.6.0+ (https://blast.ncbi.nlm.nih.gov/Blast.cgi).
Data analysis	We provide code for all analysis as well as all the input data necessary to reproduce our results. The input data is available on the S2F dedicated website at: http://www.paccanarolab.org/s2f which is described in the main manuscript. It is also available in Zenodo at DOI: 10.5281/zenodo.5514322 The code is available at https://github.com/paccanarolab/s2f which is also linked from http://www.paccanarolab.org/s2f as well as Zenodo with DOI: 10.5281/zenodo.5513070

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

We used all publicly available datasets. Input data to reproduce our results are provided at http://www.paccanarolab.org/s2f and Zenodo with DOI: 10.5281/zenodo.5514322

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Our experiments were carried out on all bacteria annotated in GOA, with at least 10 experimental annotations, involving at least 8 GO terms, and with at least 3 genes annotated in each GO subdomain (MF, BP and CC). These criteria were chosen so that each organism used for evaluating S2F had at least a few experimentally annotated genes (to be able to assess the performance in a per-gene setting) while maintaining a reasonable diversity of annotated GO terms (to be able to assess the performance in a per-term setting) in GOA. These criteria resulted in a set of 10 bacteria, whose combined sequences contained a total of 34,064 proteins.
Data exclusions	All the proteins from the 10 bacteria detailed above were used in our experiments -- no exclusions.
Replication	Code to fully reproduce the results is publicly available at https://github.com/paccanarolab/s2f which is also linked from http://www.paccanarolab.org/s2f as well as Zenodo with DOI: 10.5281/zenodo.5513070
Randomization	In order to mimic real world scenarios, we used the entire annotated protein set of a given bacteria as a test set. In other words, for each bacteria, we assume that we have no protein annotations and we check how well S2F predicts the GO terms for those proteins that have been experimentally annotated. In figures 2-5, error bars were estimated using 10,000 bootstrap iterations on the gene set or term set, respectively.
Blinding	Not applicable

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging