Bioinformatics, 36(14), 2020, 4116–4125
doi: 10.1093/bioinformatics/btaa281
Advance Access Publication Date: 30 April 2020
Original Paper

Genome analysis

Optimization of co-evolution analysis through phylogenetic profiling reveals pathway-specific signals

Idit Bloch , Dana Sherill-Rofe†, Doron Stupp†, Irene Unterman, Hodaya Beer, Elad Sharon and Yuval Tabach*

Department of Developmental Biology and Cancer Research, Institute for Medical Research Israel-Canada, Hebrew University of Jerusalem, Jerusalem 9112102, Israel

*To whom correspondence should be addressed.
†The authors wish it to be known that these authors contributed equally.

Associate Editor: Arne Elofsson

Received on November 4, 2019; revised on April 17, 2020; editorial decision on April 21, 2020; accepted on April 23, 2020

Abstract

Summary: The exponential growth in available genomic data is expected to reach full sequencing of a million genomes in the coming decade. Improving and developing methods to analyze these genomes and to reveal their utility is of major interest in a wide variety of fields, such as comparative and functional genomics, evolution and bioinformatics. Phylogenetic profiling is an established method for predicting functional interactions between proteins based on similarities in their evolutionary patterns across species. Proteins that function together (i.e. generate complexes, interact in the same pathways or improve adaptation to environmental niches) tend to show coordinated evolution across the tree of life. The normalized phylogenetic profiling (NPP) method takes into account minute changes in proteins across species to identify protein co-evolution. Despite the success of this method, it is still not clear what set of parameters is required for optimal use of co-evolution in predicting functional interactions. Moreover, it is not clear if pathway evolution or function should direct parameter choice. Here, we create a reliable and usable NPP construction pipeline. We explore the effect of parameter selection on functional interaction prediction using NPP from 1028 genomes, both separately and in various value combinations. We identify several parameter sets that optimize performance for pathways with certain biological annotation. This work reveals the importance of choosing the right parameters for optimized function prediction based on a biological context.

Availability and implementation: Source code and documentation are available on GitHub: https://github.com/iditam/CompareNPPs.

Contact: yuvaltab@ekmd.huji.ac.il

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Phylogenetic profiling (PP) was first introduced by Pellegrini (Pellegrini *et al.*, 1999) more than 20 years ago to assign protein function through evolutionary similarity. PP follows the pattern of presence or absence of protein orthologs in a set of genomes, tracing its evolutionary history (Fig. 1a). Proteins (or genes) with similar loss and retention patterns throughout the tree of life are considered co-evolved and tend to be functionally coupled. Multiple approaches have been suggested for extraction of information from phylogenetic data (Arkadir *et al.*, 2019; Date and Marcotte, 2003; Dey *et al.*, 2015; Li *et al.*, 2014, 2019; Niu *et al.*, 2017). PP has been widely used to predict protein function (Eisen and Wu, 2002; Enault *et al.*, 2004; Jiang, 2008), protein–protein interactions (Kim and Subramaniam, 2005; Sun *et al.*, 2005), protein subcellular localization (Marcotte *et al.*, 2000; Pagliarini *et al.*, 2008), cellular organelle location (Avidor-Reiss *et al.*, 2004; Hodges *et al.*, 2012) and protein annotations (Merchant *et al.*, 2007).

Methods that implement the PP approach encounter several challenges. First, considering the existence (i.e. presence or absence) of protein orthologs in an organism depends on a minimal similarity threshold. More stringent sequence similarity requirements lead to a higher loss to retention ratio, altering the resulting profile and subsequent similarity estimates (Enault *et al.*, 2004; Tabach *et al.*, 2013a, b). Using a fixed value to transform sequence similarity to a binary pattern of 1 (presence) and 0 (absence) results in loss of information on partial homology (Fig. 1b). Moreover, the similarity score does not express the phylogenetic distance between the species. The meaning of similar scores can be very different, depending whether the species compared are close or distant. For example, a sequence similarity of 60% between proteins suggests poor conservation between two closely related organisms (e.g. humans and chimpanzees), or extremely high

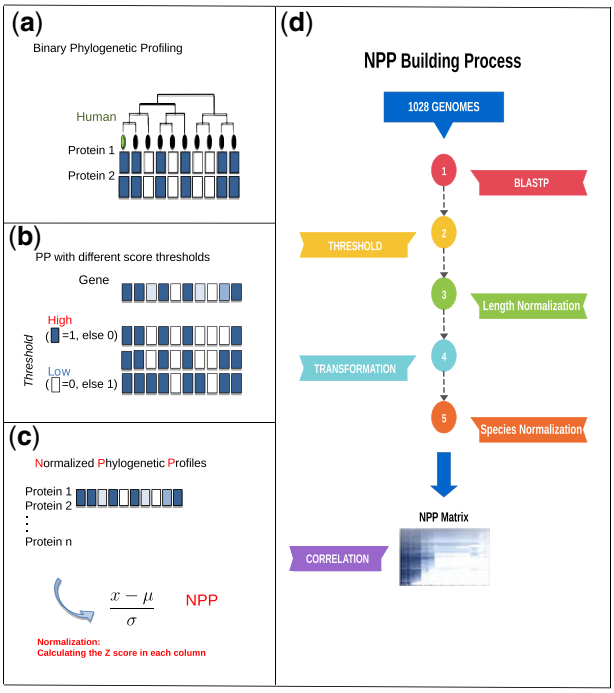

Fig. 1. NPP. (**a**) Binary PP is generated by searching for a homologous protein in a reference genome. If a homolog passes the threshold of similarity (e.g. BLASTP), the entry equals one (dark blue). For those species where no homolog passes the threshold, the entry is zero (white). (**b**) In most cases in eukaryotes, the conservation between two potential orthologous proteins is on a continuous scale between 100% conservation and complete loss. Using different thresholds results in different binary phylogenetic profiles. (**c**) The NPP uses the BLASTP score between the query protein and proteins in target genomes to calculate the similarity between the proteins in values between 0 and 1. This profile can also be represented in a graph. The numbers are then converted to *z*-scores [taking into account the evolutionary distance between the species and the reference genome (Tabach *et al.*, 2013b)]. These are used to search for similar profiles. (**d**) The process of creating the NPP table. (i) We ran BLASTP for all query genome proteins against all 1028 genomes sequences, and took the best BLASTP bit score for each protein. (ii) We set a bit score threshold value and updated all smaller bit scores to the threshold value. (iii) Each bit score was divided by the score of the query genome against itself, to normalize to the length of the query sequence. (iv) We transformed the bit score values (by log2, sqrt, 1/*x* etc.). (v) The matrix columns were normalized (*z*-score) to the evolutionary distance from the query genome. (vi) A correlation method was chosen to compare phylogenetic profiles in the NPP matrix (Color version of this figure is available at *Bioinformatics* online.)

conservation between distant organisms (e.g. humans and *Escherichia coli*). To reduce some of the biases that result from the use of thresholding and score relativity, several methods have adopted continuous conservation scores (Enault *et al.*, 2004; Franceschini *et al.*, 2016) (Fig. 1a–c). We have developed normalized phylogenetic profiling (NPP) that uses a continuous scale of conservation instead of defining a cutoff for the protein being lost or retained (Fig. 1c) (Tabach *et al.*, 2013a, b). In addition, the NPP method employs different normalizations to overcome the relative phylogenetic distance. Since 2013, we have been using our NPP method to discover genetic interactions and novel proteins in the RNA interference pathway, RNA methylation and different human diseases including cancer (Omar *et al.*, 2018; Sadreyev *et al.*, 2015; Schwartz *et al.*, 2013; Tabach *et al.*, 2013a, b). Recently, we demonstrated that NPP can be used to identify novel DNA repair factors in the homologous recombination repair pathway (Findlay *et al.*, 2018; Sherill-Rofe *et al.*, 2019). However, the NPP method is based on multiple parameters, which when adjusted may yield better performance. Despite extensive research, it is not certain that the parameters used thus far are optimal, and whether some combinations of parameters may outperform others at specific PP tasks, such as predicting different kinds of pathways.

Here, we explore these questions under the framework of the previously described NPP. In the first part of Section 3, we describe the construction of the NPP using the default parameters. The NPP matrix is based on several steps (Fig. 1d), including scoring the alignment between the reference species and all other species, converting to a conservation score and normalization. After constructing the matrix, profile similarity must be measured between protein profiles to assess functional relatedness. All steps have tunable parameters that we explore to identify both the best combinations overall as well as per-pathway combinations, thus introducing fitted implementations of NPP parameters for specific purposes.

2 Materials and methods

2.1 Building the NPP matrix

The NPP matrix is based on data including 20 192 *H.sapiens* proteins (the reference proteome) and the proteomes of 1028 eukaryotes species. To build the matrix, we downloaded the *H.sapiens* proteome as well as 1028 eukaryotes proteomes from UniProt reference proteomes (June 2018 release) (UniProt Consortium, 2018, 2019). The protein identifiers in the reference (human) proteome were converted to gene symbols. Proteins that did not map to gene symbols were omitted from the data. We then performed three filtration steps for the reference proteome - excluded proteins with sequence length <40 amino acids, filtered identical sequences and in the case of multiple proteins annotated to the same gene symbol (isoforms) we took only the protein with the longest sequence. Therefore, we use proteome and genome (and respectively, proteins and genes) interchangeably. Then BLASTP was used to score each human protein with its best match in each species. The filtered human proteome was used as 'query' and each of the 1028 species proteomes as 'subject' for the BLASTP command-line application. BLASTP command was run with '-max_target_seqs 1' to keep only the best aligned sequence for each query gene in each species. After these steps, we obtained a matrix M of 20 192 rows (genes)×1028 columns (species), where $M_{i,j}$ = bit score of protein i sequence search in species j.

2.2 Test methods

We performed two tests to compare different NPP parameters.

2.2.1 ROC AUC using CORUM complexes

For the first test, we used a CORUM complexes dataset, retrieved as an adjacency list (SIF format) from PathwayCommons (v10, https://www.pathwaycommons.org). This list contained a 'positive' set of 35 221 gene pairs and was matched with a 'negative' set of 171 106 random pairs of genes, filtered for duplicates and positives. The list of positives and negatives are attached in the Supplementary Material. To measure the predictive performance of a combination of specific parameters, the correlation between each gene pair's phylogenetic profiles was calculated and a ROC curve and ROC AUC was calculated using the R package 'ROCR' (Sing *et al.*, 2005).

2.2.2 ES scores of KEGG pathways

For the second test, we used an enrichment score (ES) in an algorithm similar to GSEA to assess the influence of different NPP parameters on the ability to predict KEGG pathways for each parameter combination. Pathways were obtained from KEGG, as of May 2018, using the 'KEGGREST' R package (Tenenbaum, 2019). Pathways with <15 genes or more than 400 genes were filtered, with 298 pathways remaining. We used a list of 203 848 336 gene pairs, which covers all possible gene pairs in the NPP. The list was ranked by correlation and the correlation transformed from range (−1, 1) to range (0, 2) by adding 1, to avoid negative values. Then for each pathway P in the dataset, we used the GSEA algorithm to calculate the ES of the gene pairs in P with regard to the ranked gene pairs list. Briefly, each gene pair in the all gene pairs list receives a score for either belonging to the given pathway (the correlation between the gene pair) or not (a constant negative value). A cumulative sum of these scores is calculated, ordered by the correlations

between all gene pairs (descending), and the maximum of this cumulative sum is used as the ES score of the pathway (Supplementary Fig. S2a).

To estimate the significance of the ES values, we first created 1000 random gene sets for each gene set size of KEGG pathways (KEGG pathway sizes vary between 15 and 352 genes, with the majority of pathways containing <100 genes). The ES can be calculated for each random set for each parameter combination. Supplementary Figure S2b shows the ES values of one of the parameter combinations for KEGG pathways along with ES values for the random gene sets. After testing the random sets for to all KEGG pathway sizes together (Supplementary Fig. S2b), we chose five representative gene set sizes: 15, 30, 50, 100 and 350, and calculated the ES of 1000 random gene sets for each of the five sizes (Supplementary Fig. S2c). These sizes cover the range of KEGG pathway sizes. As can be seen in Supplementary Figure S2b and c, the representative sizes create similar trends of distributions and give accurate estimation of the ES scores of the random sets for all sizes, which are computationally heavy. We then analyzed the distribution of the ES values of the random groups that are equal or above the mean and ignored the values that are smaller than the mean. When mirroring the distribution around the average (using only values higher than average to estimate the distribution), it becomes a normal distribution and we can calculate the SDs, the *z*-scores and the equivalent estimated *P*-values (Supplementary Fig. S2c). The *z*-score of the ES scores that derived from these calculations was used to compare between different parameter combinations for each KEGG pathway, representing the distance in units of SDs from the mean of random ES values. The mirroring slightly increases the variation of the random gene set ES distributions, therefore, reducing the *z*-scores and creating a more conservative estimation of the ES significance. The ES *z*-score is referred to as ES score throughout the text.

To compare the results of the ES score test with those of ROC AUC, we examined the distribution of the KEGG pathway scores in each of the parameter combinations. Supplementary Figure S7 demonstrates that the distributions are consistent with ROC AUC and that ES scores tend to be higher for combinations that yield higher AUC values.

2.3 The building parameters

In the single parameter analysis, when testing each of the five parameters, the other four parameters were fixed to a default value. The default values were: threshold =20.4; length normalization - yes; transformation - $\log_2(x)$; species normalization - yes; and correlation method - Pearson. These values correspond to the normalization used in previous uses of NPP (Sherill-Rofe *et al.*, 2019; Tabach *et al.*, 2013a, b).

2.3.1 Threshold

For a threshold value T, All BLAST bit scores in the matrix that were less than T, or missing, were assigned the value T.

2.3.2 Length normalization

In the length normalization stage, we divided all values in the matrix columns by the bit scores of the query (human) protein against itself. Then, we set all values >1 to be 1, in order to deal with exceptional cases where other species received bit scores that were higher than the self-hit.

2.3.3 Transformation

We tested six monotonic transformations: $\sqrt{x}$, $\log_2(x)$, $1/\sqrt{x}$, $1/x$, $1/x^2$ and $1/x^3$.

2.3.4 Species normalization

We normalized the phylogenetic distance between species by *z*-scoring the score:

$$M_{i,\ j} = \left(M_{i,\ j} - \mu_j\right)/\sigma_j$$

$M_{i,\ j}$ = score of protein i sequence search in species j
μ_j = mean of all scores of species j
σ_j = SD of all scores of species j.

2.3.5 Correlation

We tested the Pearson, Spearman and Kendall correlation methods.

2.4 Parameter combinations

We took 14 out of 19 parameter values of the 5 parameters described above and checked all their possible combinations (as described in Section 3). The 14 values are marked in bold in Table 1. All combination values are listed in Supplementary Table S1. We then tested all selected combinations for ROC AUC in CORUM complexes and ES score for KEGG pathways as described above. The results are listed in Supplementary Tables S2 and S3, respectively. The combinations heatmap (Fig. 3a) was built using the 'ComplexHeatmap' R package (Gu *et al.*, 2016). Rows and columns of the heatmap were clustered by hierarchical clustering based on Euclidean distance and the 'complete' linkage method.

To create the compressed heatmap presented in Supplementary Figure S5, we used the R function 'cutree' to cut the columns dendrogram to six sub-clusters. For each cluster, we ranked separately the ES scores and the ROC AUCs of the cluster combinations, so that each of the cluster combinations received two rank values. For each cluster, we chose the combination that yielded the maximal sum of the two ranks as representing combination for the cluster. For each of the six representative combinations, we calculated in how many pathways out of 298 KEGG pathways was the combination's ES score the highest among all six combinations, as shown in the upper bar plot in Supplementary Figure S5.

2.5 Comparison to other methods

The performance of NPP recommended general parameters combination ('generally recommended combination') was compared to the methods 'PrePhyloPro' and 'Binary hamming' as described in Section 3. The 'generally recommended combination' consists of the parameter values: threshold =20.4 after species normalization; without length normalization; transformation $=1/x$; with species normalization; and Pearson correlation. For implementing the PrePhyloPro method, we used the same BLASTP results as the NPP, taking the BLAST *E*-value instead of the BLAST bit score as protein similarity measurement and proceeded according to the PrePhyloPro algorithm (Niu et al., 2017). The 'Binary hamming' data were constructed by using the BLAST *E*-value, with a threshold $T=10^{-3}$, replacing values larger than 10^{-3} with 0 (homolog absence) and values smaller than 10^{-3} with 1 (homolog present). The hamming distance was used as a similarity metric between pairs of protein profiles.

2.6 Computational environment

All analyses were run on a high-performance computing system (Hebrew University BioCompute iCore cluster) running debian UNIX, using the SLURM job scheduling system and R language version 3.5.1. Analysis pipelines were built using the snakemake (Koster and Rahmann, 2012) workflow engine.

3 Results

3.1 The NPP matrix building parameters

The process of establishing the NPP dataset includes six steps (Fig. 1d). First, BLASTP (Camacho *et al.*, 2009) is used to score each protein in the reference proteome against all available eukaryotes species, currently encompassing 1028 eukaryotes genomes, taking the best BLASTP bit score for each ortholog in each species. We obtain a matrix of bit scores in which the rows represent proteins and the columns represent species. Some BLASTP bit score values are

Table 1. Parameters test results

Parameter	Parameter value	CORUM complexes ROC AUC	KEGG pathways ES score Median (IQR)
Threshold	**No threshold**	0.686	8.3 (5.56–13.2) P-value$<10^{-17}$
	T =20.4	0.68	6.48 (4.04–10.8) P-value$<10^{-11}$
	T=40	0.674	5.97 (3.45–10.2) P-value$<10^{-9}$
	T=20.4 after species Normalization	0.672	8.52 (5.49–13.7) P-value$<10^{-18}$
	T=0.01 after length normalization	0.679	7.21 (4.59–12.2) P-value$<10^{-13}$
	T=0.05 after length normalization	0.657	6.52 (3.75–10.3) P-value$<10^{-11}$
Length normalization	**With**	0.68	6.48 (4.04–10.8) P-value$<10^{-11}$
	Without	0.684	6.32 (3.87–9.65) P-value$<10^{-10}$
Transformation	No transformation	0.633	4.67 (2.7–9.52) P-value$<10^{-6}$
	$\sqrt{x}$	0.658	5.79 (3.51–10.3) P-value$<10^{-19}$
	$\log_2(x)$	0.68	6.48 (4.04–10.8) P-value$<10^{-11}$
	$1/\sqrt{x}$	0.694	5.6 (3.4–9.32) P-value$<10^{-8}$
	$1/x$	0.693	4.5 (2.38–6.64) P-value$<10^{-6}$
	$1/x^2$	0.642	1.59 (0.18–3.17) P-value$<10^{-2}$
	$1/x^3$	0.602	1.56 (0.24–3.23) P-value$<10^{-2}$
Species Normalization	**With**	0.68	6.48 (4.04–10.8) P-value$<10^{-11}$
	Without	0.546	2.37 (−0.11–4.96) P-value$<10^{-3}$
Correlation method	**Pearson**	0.68	6.48 (4.04–10.8) P-value$<10^{-11}$
	Spearman	0.673	4.66 (2.69–7.81) P-value$<10^{-6}$
	Kendall-tau	0.672	3.94 (2.39–7.19) P-value$<10^{-5}$

Note: A summary of the results of the two tests, ROC AUC and ES score, for each of five parameter values (see text). The parameter values that were taken for the combinations analysis are indicated in bold. The P-values refer to the ES score median values.

very low or missing (for sequences with no similar sequence in the subject genome). Second, we set a thresholding parameter for bit scores. Low or missing bit scores indicate the lack of significant homolog for the protein or that the homolog was dramatically drifted throughout evolution. These scores might cause artifacts in the correlations found between protein conservation patterns and lead to inaccurate prediction of protein functional relations. To reduce this unwanted influence, we set a bit score threshold value T, which is the minimal bit score value across all species that corresponds to BLAST E-values $\leq$0.05, and assign all missing or low bit scores ($<T$) to T (Tabach *et al.*, 2013a, b). Third, we normalize bit scores to the bit score of the self-hit (termed length normalization). Bit scores are affected by the length and amino-acid content of the reference organism proteins. Thus, to normalize to the protein length, we divide each bit score by the score of the reference protein against itself, resulting in values between 0 and 1 (Enault *et al.*, 2004). Fourth, we tackle the difference in the length-normalized score distributions among different species. In species that are more distant from the reference genome, it is less common to find proteins with high similarity to the reference genome protein, thus most of the similarity scores between distance species are closer to 0 and their distributions tend to be skewed to the left. The greater the phylogenetic distance, the stronger the skewness (Supplementary Fig. S1a). Conversely, most values in species that are closely related to the reference genome are close to 1. In order to make the distributions symmetric and apt to comparison, and considering the fact that the majority of species in the NPP are distant from the reference organism, we add a monotonic transformation stage, using log2 (Tabach *et al.*, 2013a, b). The log2 transforms right-tailed distributions to be more symmetric around the mean (Supplementary Fig. S1b) and improves the ability to compare protein conservation among different species. Fifth, we perform standard scaling of the values (z-scores) in each species. This is done in order to identify proteins that are more or less conserved than expected, based on the evolutionary distance from the reference organism. The outcome is a matrix that describes the conservation pattern of the proteins of the reference species across eukaryotes and considers minute changes in protein evolution. The NPP matrix is used to identify functional interactions among proteins, based on the similarity between their phylogenetic profiles. Sixth, we use Pearson correlation to identify co-evolution between protein pairs. Each of these steps can be done using different parameters, which when adjusted may yield better performance.

3.2 Evaluation metrics

To compare the different parameters, two methods were used - area under the curve of the receiver operating characteristic curve (ROC AUC) and enrichment score (ES).

ROC AUC was calculated based on the CORUM database of manually annotated protein complexes from mammalian organisms (Giurgiu *et al.*, 2019), similar to previous work (Niu *et al.*, 2017). The CORUM database contains 35 221 pairs of proteins that participate in shared complexes. This database was chosen because protein complexes have been shown to be frequently co-evolved (Dey and Meyer, 2015; Li *et al.*, 2014). In the ROC AUC, we compared the effect of different parameters on NPP ability to predict co-occurrence of two proteins in a complex in terms of specificity and sensitivity, using all 35 221 pairs from the CORUM database as positives and random protein pairs that do not belong to the same complex as negatives in a 1:5 positive to negative ratio.

For the second metric of ES, we used the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways database, which contains over 300 pathways (Kanehisa, 2019; Kanehisa and Goto, 2000). KEGG is a widely used resource of curated knowledge from scientific literature on biological molecular interactions and reaction networks. For each pathway in KEGG pathways, we calculated an ES based on the GSEA algorithm (Subramanian *et al.*, 2005), to assess how NPP with different parameters assigns proteins to the pathway. The algorithm uses a list of all protein pair correlations according to the NPP, ranked from highest to lowest correlation. The ES reflects the degree to which the pathway is over-represented at the top of the ranked list. The higher the ES, the more the pathway is represented at the top of the list, where the higher correlation values are present, and the better NPP is at assigning proteins to the pathway. The ES method quantifies the ability of NPP to identify whole pathways. To assess the significance of the score, we compared the ES of each pathway with the scores of 1000 random protein sets of the same size as the tested pathway in terms of z-scoring (Supplementary Fig. S2), which is equivalent to the estimated P-value of the ES. These z-scores, calculated for each parameter combination independently, were used to compare between different NPP parameters.

Using the two resources (CORUM complexes & KEGG pathways) provides us with a full picture of the ability of NPP with the different parameters to identify protein interactions. ROC analysis quantifies the ability of the method to identify protein pairs belonging to a common complex, while ES gives a score to all proteins in

each pathway and assesses the ability of NPP to identify whole pathways.

3.3 Single parameter analysis

In each of the NPP analysis steps, except for orthologs detection, which was always based on BLASTP, we tested multiple parameter values while keeping all other parameters constant. We did not compare BLASTP to other methods of orthologs scoring that identify very low similarity orthologs because very low scores were removed through thresholding and were not further considered in the analysis. We used the two tests described above, ROC AUC and ES score, to test the parameters (Table 1). The analysis was performed using *H. sapiens* as the reference organism.

3.3.1 Threshold

We began the analysis by assessing the effect of different threshold values. We tested six methods of setting a threshold value. First we tested T=20.4, which was the minimal bit score value across all species that corresponds to BLAST E-values $\leq$0.05. We also tested T=40 and T=0 to see the influence of a higher threshold and of no threshold at all, respectively. In addition, we tested postponing the threshold stage past the length normalization stage, when the values are between 0 and 1, and then replacing values $<T$' with T', where T' is 0.01 or 0.05. We also tested thresholding after the species normalization step with a threshold of T=20.4, i.e. these values are omitted from the length normalization, transformation and species normalization steps, and assigned the lowest value per column, received at the very last stage. The CORUM (ROC AUC) analysis showed small differences in the AUC among the different threshold values. On the other hand, threshold 20.4 after species normalization performed best under the KEGG pathways analysis (ES). Late flooring and threshold 40 seem to have the lowest overall performance. For the combination analysis, we further examined the three best parameter values according to each test: no threshold, threshold = 20.4, threshold = 0.01 after length normalization and threshold = 20.4 after species normalization. These four options were taken for further analysis in the parameter combinations stage, to examine their influence when combined with other parameter values (Table 1, Fig. 2a and b).

3.3.2 Length normalization

We tested the NPP with and without length normalization (dividing each bit score by the score of the reference protein against itself). Both tests showed that performing length normalization does not significantly boost performance (Table 1, Supplementary Fig. S3a and b). We took both options, with and without length normalization, for the parameter combinations analysis.

3.3.3 Transformation

For the NPP building step that handles the skewness of the score distributions, we tested six monotonic transformations of log and negative powers that are known to adjust right-tail distributions and transform them to more symmetric distributions. These are common Box–Cox power transformations (ordered from weakest to strongest): $\sqrt{x}$, $\log_2(x)$, $1/\sqrt{x}$, $1/x$, $1/x^2$ and $1/x^3$. The stronger the transformation, the more the values are condensed toward the mean, reducing the right-tail that results from few outlier proteins that have homologs with high bit score within a species. We also tested not performing a transformation at all. Both tests showed that omitting the transformation stage yields poor results (Supplementary Fig. S3c and d) and so are the results when using the $1/x^2$ and $1/x^3$ transformations, whereas $\log_2(x)$ and $1/\sqrt{x}$ were among the three best parameter values in both tests (Table 1, Fig. 2c and d). There were differences in the results of $\sqrt{x}$ and $1/x$ between the two test methods, as can be seen in Table 1. We decided to take the two best results by ROC AUC (the transformations $1/x$ and $1/\sqrt{x}$) and the best result by ES score [$\log_2(x)$] for the parameter combinations analysis. We did not take the second best result by ES score ($\sqrt{x}$), as it had a relatively low ROC AUC value.

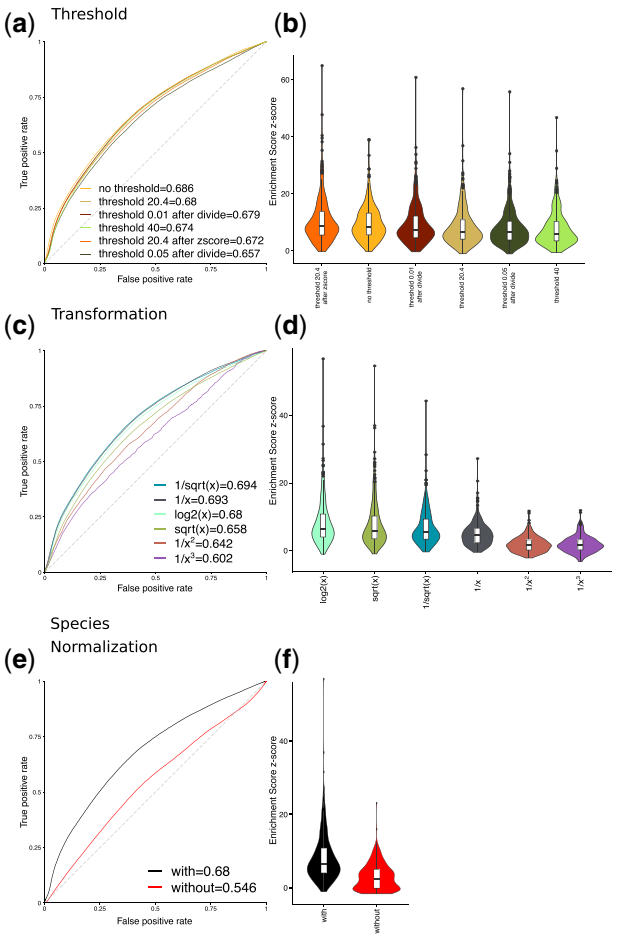

Fig. 2. Parameters test results. The results of the two tests for three parameters, which had substantial differences among different parameter values. The results are also summarized in Table 1. (**a**, **c**, **e**) The ROC curve analysis of NPP built with different parameters. NPP correlations for protein pairs from CORUM (positive) were compared to control (negative). The X and Y axes represent the false-positive rate (FPR=1 − specificity) and true-positive rate (TPR, sensitivity), respectively. (**b**, **d**, **f**) The distributions of the ES score values of NPP built with different parameters for 298 KEGG pathways. (**a**, **b**) The effect of thresholding on NPP, (**c**, **d**) transformation and (**e**, **f**) species normalization

3.3.4 Species normalization

We tested the NPP with and without normalizing to the overall species distance from humans. Both tests showed that species normalization markedly improves performance (Table 1, Fig. 2e and f). Nonetheless, we proceeded with both options, with and without species normalization, for the combination analysis, to elucidate any synergistic effect with other parameter values.

3.3.5 Correlation

Using the NPP matrix to compare phylogenetic profiles requires a similarity metric. We tested three common correlation methods that are suitable for continuous data, such as NPP values: Pearson correlation coefficient, Spearman's rank correlation coefficient and Kendall's tau rank distance. We did not test Euclidean distance as it is equivalent to Pearson correlation when applied to row-normalized profiles. Both tests demonstrated the relative strength of Pearson correlation. However, the difference among the test results for the different correlation methods was small; therefore, we proceeded with all three for further analysis (Table 1, Supplementary Fig. S3e and f).

For each parameter, we chose the best performing values, and moved on to explore all the combinations, to check if better results

Fig. 3. NPP parameter combinations heatmap. (**a**) Heatmap of the ES scores of 298 KEGG pathways in 126 different combinations of the NPP parameters. Every row in the heatmap represents one pathway, and every column represents one combination of parameters. The additional columns on the left represent the PrePhyloPro and the binary methods. The annotation on the left indicates the classification of the pathways to six main classes, according to the KEGG database. The ROC AUC line plot on the top shows the AUC for each combination while the median ES score line plot shows the median value of each combination over all pathways. The ROC AUC and the median ES score values for the PrePhyloPro and the binary methods are marked in orange lines on the left side of the line plots. Three rectangles marked a–c are discussed in Supplementary Figure S3. The columns' dendrogram is colored by the division into six clusters, which are annotated by numbers 1–6. (**b**) For each class in the KEGG pathway classifications, the distribution of the median ES score is presented. The median is calculated for each pathway in the class (rows), over all tested combinations (columns)

could be obtained both in general and for prediction of different pathways.

3.4 Parameter combinations

To find the optimal parameter set, we tested all combinations of 14 out of 19 parameter values (marked in bold in Table 1) and removed the parameters that dramatically reduce performance. We applied the same two analyses, based on the CORUM and KEGG databases, to evaluate performance. The results of the ROC AUC test on all combinations are presented in Supplementary Table S1. Figure 3a contains ES score test results for all combinations. This heatmap shows the ES scores of 298 KEGG pathways.

3.4.1 Co-evolution in different pathways is better captured by using different parameter sets

As shown in Figure 3a, although some combinations are better overall, the performance results show a complex pattern. Species normalization and thresholding had the largest impact of performance. Species normalization had an overall positive effect, as it highlights values that differ from the expected conservation. Using Pearson correlation improves the results in most of the combinations. Inversely, no thresholding sometimes resulted in a reduced prediction of interactions within pathways due to the added noise in the low values. Also, different combinations are better at identifying different pathways. For example, we found that pathways that are classified by KEGG as 'Environmental Information Processing' received very high ES scores, while metabolism pathways had relatively low scores in most of the NPP combinations (Fig. 3b).

Focusing on three areas in the heatmap, marked in Figure 3a (Supplementary Fig. S4), we found several pathways that are known to be captured well by PP, e.g. the TCA cycle (Tabach *et al.*, 2013a) and glycolysis gluconeogenesis (Dey and Meyer, 2015; Li *et al.*, 2014). Here, we show that these same pathways receive very high scores for specific combinations (rectangle 'a') compared to low scores for other combinations (rectangle 'b'). Similarly, the pathways in rectangle 'c' are related to signaling and to basic cellular processing, which are usually not captured well by PP (Dey and Meyer, 2015). These pathways benefit from a different set of parameters (e.g. no length normalization), where they score well.

In order to investigate the connection between pathways and parameter combinations further, we selected six combinations by clustering the parameter combinations to six clusters (marked by colored dendrogram in Fig. 3a) and choosing for each cluster the best combination among the cluster combinations. The scores of these six combinations are shown in Supplementary Figure S5 with the list of all included pathways. Combination clusters 1 and 3 most frequently obtained the best ES scores among all KEGG pathways. Table 2 lists the test results of the six representing combinations.

3.4.2 Recommended parameters

Overall, we show that the choice of parameters has a major effect on our ability to predict proteins functional interactions. When using the NPP in general analyses, we recommend using the parameters values of combination 3 in Table 2 that received high scores in both tests. The parameters in this combination are: threshold-20.4 after species normalization; length normalization - no; transformation - $1/x$; species - yes; and correlation method - Pearson. In cases when one type of pathway is of interest we recommend choosing suitable parameters according to Supplementary Table S3.

3.5 Comparison to other methods

We compared the p bfvdzm erformance of our method with two established PP methods - the 'PrePhyloPro' (Niu *et al.*, 2017) and a binary PP that uses Hamming distance as the similarity metric between profiles ('Binary Hamming') (Kensche *et al.*, 2008; Niu *et al.*, 2017). Recently, Niu *et al.* (2017) demonstrated that their PP method 'PrePhyloPro' achieved the best performance followed by NPP. As such, we decided to compare our performance with 'PrePhyloPro' in general and for every KEGG pathway. We applied both evaluation methods of the ROC AUC and the KEGG pathway ES on 'PrePhyloPro' and 'Binary Hamming' methods. The ROC AUC results were 0.672 for 'PrePhyloPro' and 0.669 for 'Binary hamming'. Both AUC were lower than the results of the NPP with the recommended general parameter values ('generally recommended combination'), which yielded an AUC of 0.698 (Supplementary Fig. S6a). NPP with other parameter combinations received even higher AUC values like 0.702 (see Supplementary Table S2).

Even better results were shown in the second analysis of the KEGG pathways (ES). The ES score showed better performance to NPP 'generally recommended combination' in 56% of the pathways comparing to 'PrePhyloPro' and in 79% of the pathways comparing to 'Binary Hamming' (Supplementary Fig. S6b and d). The median (IQR) of the ES scores of all pathways was 8.42 (5.31, 12.69) for the generally recommended combination NPP, while 'PrePhyloPro' and 'Binary Hamming' results were median (IQR) of 7.79 (5.17, 11.4) and 5.4 (3.3, 8.66), respectively.

Overall, we showed that when optimizing the parameter combinations of the NPP method, it outperforms these state-of-art PP methods.

Another method that is designated to predict protein interactions is 'mirrorTree' (Pazos and Valencia, 2001), that quantifies the similarity between phylogenetic trees of proteins, represented by their distance matrices. The computational complexity of 'mirrorTree' and its various variations and implications (Juan *et al.*, 2008; Ochoa *et al.*, 2015) make it hard to implement for the large scale of the genomes we used in the NPP, therefore we did not compare our results to this method.

Table 2. Test results of six representing combinations

Combination Cluster	Combination parameters	CORUM complexes ROC AUC	ES z-score of KEGG pathways—median (IQR)	Best in # pathways
1	T=20.4; without length normalization; transformation =$1/x$; with species normalization; Pearson correlation	0.688	8.83 (5.36–13.96) P-value<10^{-19}	134
2	T=0.01 after length normalization; with length normalization; transformation =$1/\sqrt{x}$; with species normalization; Pearson correlation	0.689	6.55 (4.2–10) P-value<10^{-11}	27
3	T=20.4 after the species normalization step; without length normalization; transformation =$1/x$; with species normalization; Pearson correlation	0.698	8.42 (5.31–12.69) P-value<10^{-17}	94
4	T=20.4 after the species normalization step; with length normalization; transformation =$1/\sqrt{x}$; with species normalization; Pearson correlation	0.699	5.43 (3.17–8.62) P-value<10^{-8}	18
5	No threshold; without length normalization; transformation =$1/\sqrt{x}$; with species normalization; Kendall-tau correlation	0.609	3.44 (1.99–5.28) P-value<10^{-4}	10
6	No threshold; with length normalization; transformation =$\log_2(x)$; without species normalization; Kendall-tau correlation	0.589	3.97 (1.73–7.03) P-value<10^{-5}	15

Note: Results of ROC AUC and ES tests for the six best parameter combinations. Listed for each combination is the number of clusters in the compressed combination heatmap (Supplementary Fig. S5), the parameter values, the AUC result of CORUM ROC test, the median and IQR of the ES scores over all KEGG pathways with the corresponding P-value of the median and the number of KEGG pathways in which the combination received the best ES score.

Combination that outperforms all combinations in most of the pathways.

Generally recommended combination.

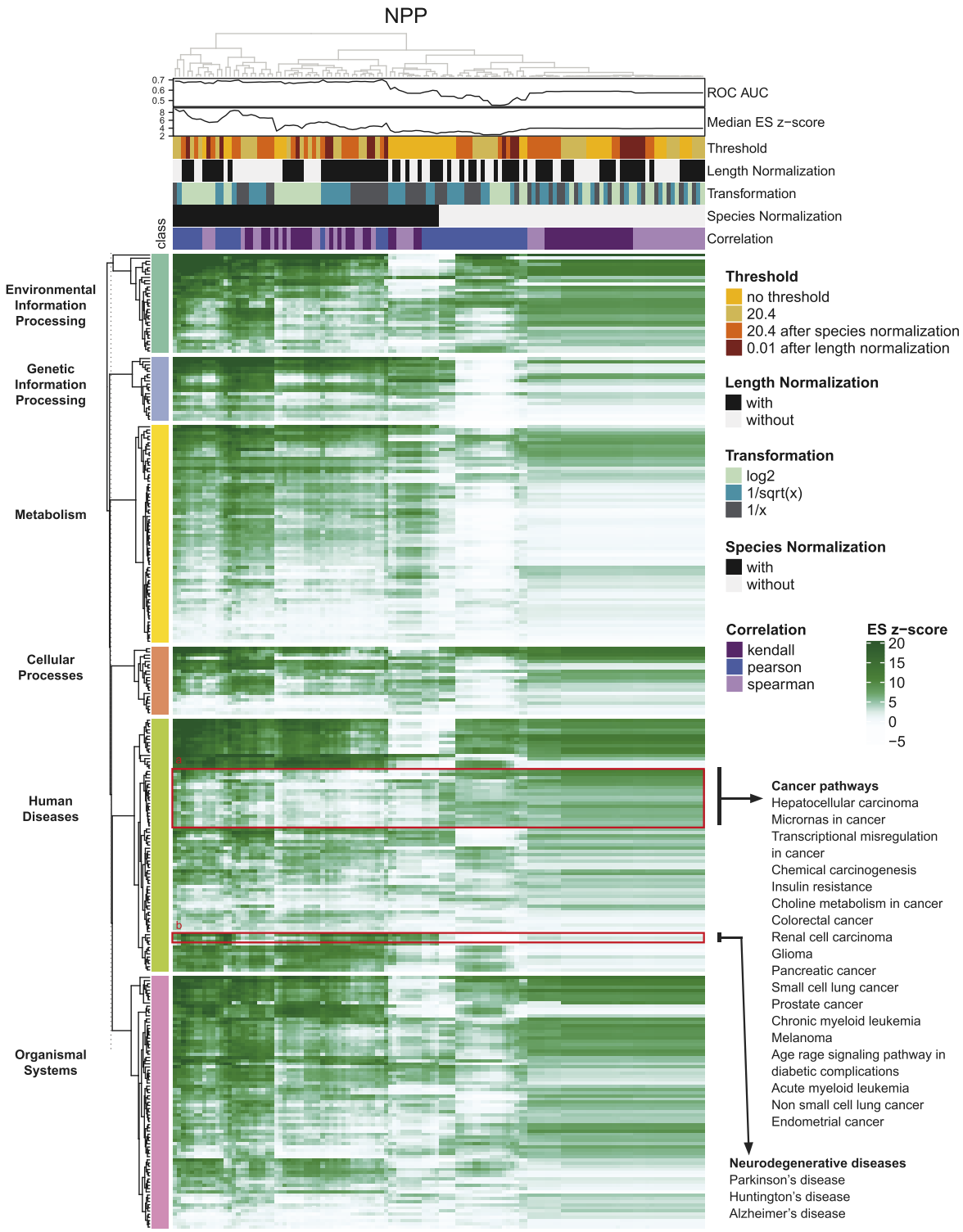

Fig. 4. NPP parameter combination heatmap - by pathway classes. Heatmap of the ES scores of 298 KEGG pathways in 126 different combinations of the NPP parameters. Rows and columns represent pathways and parameter combinations, respectively. The heatmap rows are clustered and split according to the KEGG classifications of the pathways into six classes. A group of cancer-related pathways, clustered together, is marked in rectangle 'a'. Rectangle 'b' marks a cluster of pathways associated with neurodegenerative diseases

4 Discussion

We aimed to improve the utilization of NPP to detect functional interactions between proteins. In the first stage of the analysis, we examined the effect of every NPP parameter separately on NPP ability to discern functionally linked protein pairs from random pairs, and in the second stage, we analyzed combinations based on the best parameters for each step. We showed that when the parameter combination is optimized, NPP outperforms both 'PrePhyloPro' and 'Binary Hamming' PP methods.

4.1 Single parameter analysis

We first looked at the effect of thresholding and found little difference among different threshold values, even when using no threshold at all. There was also no major difference in both ROC AUC and ES score tests when the length normalization step was omitted. This result was surprising given previous works that suggested length normalization improves prediction performance (Enault *et al.*, 2004; Tabach *et al.*, 2013b). The transformations $1/x$, $\log_2(x)$ and $1/\sqrt{x}$ performed better than other transformations, such as $1/x^2$ and $1/x^3$, aimed at increasing symmetry of the NPP value distributions (Supplementary Fig. S1). It is reasonable to conclude that NPP requires a transformation which is not too strong, that will preserve the original variation of the score values and reflect the varying conservation of different proteins within each species. We then looked at species normalization, and found that the difference between the results with and without species normalization was substantial, and demonstrated the necessity of this step in the construction of the NPP matrix. These results are in line with previously published works (Enault *et al.*, 2004; Tabach *et al.*, 2013b). Finally, we compared correlation methods. Both tests showed that the Pearson correlation method was better suited to identifying functionally related pairs, as compared to Spearman and Kendall correlations.

4.2 Parameter combinations

The ES score calculation for all selected parameter combinations showed that the largest influence was performing species normalization. Also, different pathways needed different parameters to be identified (Fig. 3a). Zooming in on specific pathways, we showed that pathways known to be identified by the NPP can be missed with some parameter sets. Conversely, pathways difficult to detect, such as signaling pathways, can be captured given the right combination of parameters (Fig. 3a, Supplementary Fig. S4).

Taking a closer look at pathways that are classified by KEGG as human diseases related pathways, a pattern emerges (Fig. 4). Related pathways tend to cluster together, i.e. their detection by the different parameter combinations is similar. For example, a large cluster containing multiple cancer-related pathways is marked in Figure 4 in rectangle 'a'. Interestingly, species normalization is less recommended for the detection of these pathways. On the other hand, three neurodegenerative disease pathways appear in rectangle 'b'. Here, both species normalization and length normalization dramatically contribute to the detection of these pathways. One possible explanation is that cancer-related genes tend to be conserved throughout the evolutionary tree, while genes related to neuronal pathways evolved later and are only conserved in our close relative species.

5 Conclusion

Overall, our analysis revealed that choosing the parameters for the NPP analysis can affect our ability to detect co-evolution. While the effect of many parameters might be small, there are certain parameters that have major impacts on performance. This is especially important for specific pathways. Most importantly, our analysis showed that in PP, parameter optimization depends on the biological context. Following the exponential growth in the volume of genomic data, we believe that PP analysis will become increasingly popular and could be used to identify the function of more proteins in different species. Correctly selecting the set of parameters can be very useful to optimize the function prediction of proteins in specific pathways of interest.

Authors' contributions

I.B., D.S.R., D.S. and Y.T. conceived and designed the study. I.B. developed the computational analysis and the bioinformatics framework. I.B., D.S.R., D.S., I.U. and Y.T. helped draft the manuscript. D.S., H.B. and E.S. participated in the processing of the data. All authors read and approved the final manuscript.

Funding

This work was supported by the Israel Science Foundation [grant number 1985/13]; and the Israel Cancer Association [grant number 0394837].

Conflict of Interest: none declared.

References

Arkadir,D. *et al.* (2019) MYORG is associated with recessive primary familial brain calcification. *Ann. Clin. Transl. Neurol.*, **6**, 106–113.
Avidor-Reiss,T. *et al.* (2004) Decoding cilia function: defining specialized genes required for compartmentalized cilia biogenesis. *Cell*, **117**, 527–539.
Camacho,C. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
Date,S.V. and Marcotte,E.M. *et al* (2003) Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat. Biotechnol.*, **21**, 1055–1062.
Dey,G. *et al.* (2015) Systematic discovery of human gene function and principles of modular organization through phylogenetic profiling. *Cell Rep.*, **10**, 993–1006.
Dey,G. and Meyer,T. (2015) Phylogenetic profiling for probing the modular architecture of the human genome. *Cell Syst.*, **1**, 106–115.
Eisen,J.A. and Wu,E.M. (2002) Phylogenetic analysis and gene functional predictions: phylogenomics in action. *Theor. Popul. Biol.*, **61**, 481–487.
Enault,F. *et al.* (2004) Phydbac2: improved inference of gene function using interactive phylogenomic profiling and chromosomal location analysis. *Nucleic Acids Res.*, **32**, W336–W339.
Findlay,S. *et al.* (2018) SHLD 2/FAM 35A co-operates with REV 7 to coordinate DNA double-strand break repair pathway choice. *EMBO J.*, **37**,
Franceschini,A. *et al.* (2016) SVD-phy: improved prediction of protein functional associations through singular value decomposition of phylogenetic profiles. *Bioinformatics*, **32**, 1085–1087.
Giurgiu,M. *et al.* (2019) CORUM: the comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Res.*, **47**, D559–D563.
Gu,Z. *et al.* (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, **32**, 2847–2849.
Hodges,M.E. *et al.* (2012) The evolution of land plant cilia. *New Phytol.*, **195**, 526–540.
Jiang,Z. (2008) Protein function predictions based on the phylogenetic profile method. *Crit. Rev. Biotechnol.*, **28**, 233–238.
Juan,D. *et al.* (2008) High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc. Natl. Acad. Sci. USA*, **105**, 934–939.
Kanehisa,M. (2019) Toward understanding the origin and evolution of cellular organisms. *Protein Sci.*, **28**, 1947–1951.
Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
Kensche,P.R. *et al.* (2008) Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. *J. R. Soc. Interface*, **5**, 151–170.
Kim,Y. and Subramaniam,S. (2005) Locally defined protein phylogenetic profiles reveal previously missed protein interactions and functional relationships. *Proteins*, **62**, 1115–1124.
Koster,J. and Rahmann,S. (2012) Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.
Li,Y. *et al.* (2019) Bayesian hidden Markov tree models for clustering genes with shared evolutionary history. *Ann. Appl. Stat.*, **13**, 606–637.
Li,Y. *et al.* (2014) Expansion of biological pathways based on evolutionary inference. *Cell*, **158**, 213–225.
Marcotte,E.M. *et al.* (2000) Localizing proteins in the cell from their phylogenetic profiles. *Proc. Natl. Acad. Sci. USA*, **97**, 12115–12120.
Merchant,S.S. *et al.* (2007) The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science*, **318**, 245–250.
Niu,Y. *et al.* (2017) PrePhyloPro: phylogenetic profile-based prediction of whole proteome linkages. *PeerJ*, **5**, e3712.
Ochoa,D. *et al.* (2015) Detection of significant protein coevolution. *Bioinformatics*, **31**, 2166–2173.
Omar,I. *et al.* (2018) Schlafen2 mutation in mice causes an osteopetrotic phenotype due to a decrease in the number of osteoclast progenitors. *Sci. Rep.*, **8**, 13005.
Pagliarini,D.J. *et al.* (2008) A mitochondrial protein compendium elucidates complex I disease biology. *Cell*, **134**, 112–123.
Pazos,F. and Valencia,A. (2001) Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Eng. Des. Sel.*, **14**, 609–614.
Pellegrini,M. *et al.* (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA*, **96**, 4285–4288.

Sadreyev,I.R. *et al*. (2015) PhyloGene server for identification and visualization of co-evolving proteins using normalized phylogenetic profiles. *Nucleic Acids Res*., **43**, W154–W159.

Schwartz,S. *et al*. (2013) High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. *Cell*, **155**, 1409–1421.

Sherill-Rofe,D. *et al*. (2019) Mapping global and local coevolution across 600 species to identify novel homologous recombination repair genes. *Genome Res*., **29**, 439–448.

Sing,T. *et al*. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.

Sun,J. *et al*. (2005) Refined phylogenetic profiles method for predicting protein-protein interactions. *Bioinformatics*, **21**, 3409–3415.

Tabach,Y. *et al*. (2013a) Human disease locus discovery and mapping to molecular pathways through phylogenetic profiling. *Mol. Syst. Biol*., **9**, 692.

Tabach,Y. *et al*. (2013b) Identification of small RNA pathway genes using patterns of phylogenetic conservation and divergence. *Nature*, **493**, 694–698.

Tenenbaum,D. (2019) *KEGGREST: Client-side REST Access to KEGG*. R package version 1.26.1.

UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*., **47**, D506–D515.

UniProt Consortium (2018) UniProt: the universal protein knowledgebase. *Nucleic Acids Res*., **46**, 2699–2699.