Rank-based interolog mapping for predicting protein-protein interactions between genomes

Yu-Shu Lo
Institute of Bioinformatics and Systems Biology, National Chiao Tung
Hsinchu, Taiwan
fantasyhwind@gmail.com

Chun-Chen Chen
Institute of Bioinformatics and Systems Biology, National Chiao Tung
Hsinchu, Taiwan
dream.chun@msa.hinet.net

Kai-Cheng Hsu
Institute of Bioinformatics and Systems Biology, National Chiao Tung
Hsinchu, Taiwan
piki.bi96g@g2.nctu.edu.tw

Jinn-Moon Yang
Institute of Bioinformatics and Systems Biology, National Chiao Tung
Hsinchu, Taiwan
moon@faculty.nctu.edu.tw

***Abstract*—As rapidly increasing number of sequenced genomes, the methods for predicting protein-protein interactions (PPIs) from one organism to another is becoming important. Best-match and generalized interolog mapping methods have been proposed for predicting (PPIs). However, best-match mapping method suffers from low coverage of the total interactome, because of using only best matches. Generalized interolog mapping method may predict unreliable interologs at a certain similarity cutoff, because of the homologs differed in subcellular compartment, biological process, or function from the query protein. Here, we propose a new "rank-based interolog mapping" method, which uses the pairs of proteins with high sequence similarity (E-value<10^{-10}) and ranked by BLASTP E-value in all possible homologs to predict interologs. First, we evaluated "rank-based interolog mapping" on predicting the PPIs in yeast. The accuracy at selecting top 5 and top 10 homologs are 0.17, and 0.12, respectively, and our method outperformed generalized interolog mapping method (accuracy=0.04) with the joint E-value<10^{-70}. Furthermore, our method was used to predict PPIs in four organisms, including worm, fly, mouse, and human. In addition, we used Gene Ontology (GO) terms to analyzed PPIs, which reflect cellular component, biological process, and molecular function, inferred by rank-based mapping method. Our rank-based mapping method can predict more reliable interactions under a given percentage of false positives than the best-match and generalized interolog mapping methods. We believe that the rank-based mapping method is useful method for predicting PPIs in a genome-wide scale.**

Keywords—Rank-based strategy; interolog mapping

I. INTRODUCTION

Protein-protein interactions play an essential role in cellular functions. For rapidly increasing of sequenced genomes, it has been of significant value to provide the approaches of predicting protein-protein interactions from one organism (with abundant known interactions) to another organism (with less interaction data). In other words, to reliably transfer protein-protein interaction annotation from one organism to another [1].

In recent years, the large number of protein-protein interactions, generated by large-scale experimental methods [2-4], computational methods [5-11], and integrated approaches [12, 13], have been collected in databases (e.g. MIPS [14], DIP [15], and IntAct [16]). The increasing protein-protein interactions provide an opportunity and challenges for understanding and predicting protein-protein interactions across multiple organisms. Many approaches, such as the sequence-based methods (e.g. PathBlast [17, 18]), interologs [1, 7], and structure-based methods [9, 19] have been developed.

The concept of "interologs" means: If interacting proteins A and B in one organism (source) have interacting orthologs A' and B' in another organism (target), the pair of A-B and A'-B' are called interologs. Operationally, the ortholog of a protein is defined as its best-matching homolog in another organism. Matthews et al. [7] proposed a "best-match mapping" method to predict *C. elegans* (worm) interactions from *S. cerevisiae* (yeast) interactome. This method considered all pairs of best-matching homologs (BLASTP *E*-value < 10^{-10}) of interacting yeast proteins as potential interologs.

Additionally, Yu et al. [1] extended and assessed the concept of interologs to provide a "generalized interolog mapping" method. The mapping method regards all pairs of homologs, which have joint similarities (see Methods) larger than a certain cutoff, as possible interologs. Their results showed that interaction annotation could be reliably transferred between two organisms if a pair of proteins has a joint *E*-value (J_E) $\leq 10^{-70}$.

There are interesting questions in best-match and generalized interolog mapping methods. First, best-match mapping method suffers from low coverage of the total interactome [1], because of using only best matches. For this question, Yu et al. [1] proposed the method of generalized interolog mapping. Second, in generalized interolog mapping method, the homologs of a query protein selected at a certain *E*-value would sometimes be different in subcellular location, biological process, or function from the query protein.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

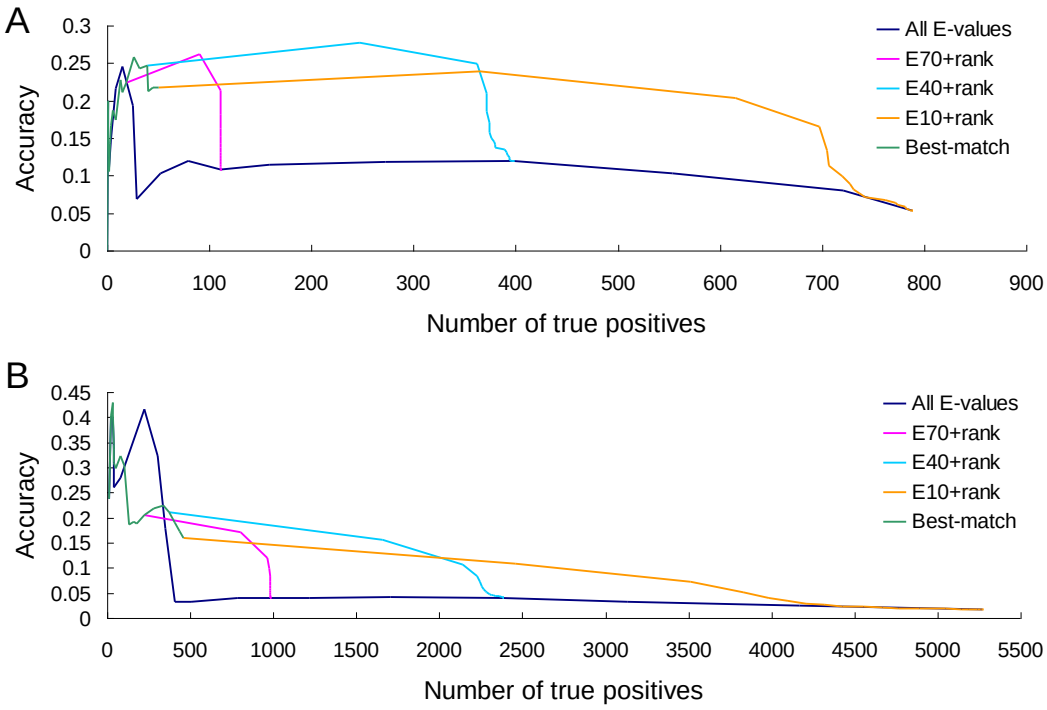

Fig. 1. The comparison of accuracy of rank-based interolog (yellow, blue, and pink), best-match (green), and generalized interolog mapping (deep blue) methods. "E10+rank", "E40+rank", and "E70+rank" mean Acc(10-10,R), Acc(10-40,R), and Acc(10-70,R), R ∈ [1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85,90,95,100, 'All'], respectively (see Methods).

For example, YLL034C in yeast has a low *E*-value (< 10^{-120}) with protein Q01853 in *M. musculus* (mouse). Q01853 is an ATPase in endoplasmic reticulum (ER), nuclear membrane and cytosol for retrotranslocation of ubiquitinated proteins from the ER into the cytosol for degradation by the proteasome [20], but YLL034C does not participates in protein degradation. The yeast protein is required for biogenesis and nuclear export of 60S ribosomal subunits. YLL034C distributes between the nucleolus, nucleoplasm, and nuclear periphery depending on growth conditions [21]. The protein pairs having these sequences may be not reliable candidates of interologs.

Orthology of different organisms are usually used in predicting interactions [22]. The third question is that, orthologous protein interactions between two species have various J_E. For example, two protein pairs P47857-P12382 and Q8R317-P40142 in mouse have orthologous interactions YMR205C-YGR240C and YMR276W-YBR117C with J_E = 10^{-171} and 10^{-27} in yeast, respectively. In other words, a certain cutoff would usually lose part of orthologous interactions.

To study these three questions, we propose a new "ranked-based interolog mapping" method for predicting protein-protein interactions between species. This method considers the homologs with higher similarity in all possible homologs (*E*-value $\leq$ 10^{-10}) as candidates of interacting proteins to gather interologs.

II. Results

A. Accuracy of rank-based interolog mapping

For practicability of approach to predict interactions, we develop a method which has reliable predicting accuracy and acceptable coverage of the interactome. In this paper, we propose a new "rank-based interolog mapping" method. This method looses the best-match mapping to get a higher coverage of the total interactome. On the other hand, this method selects part, not all, of homologs in the target organism to amend the two questions of generalized interolog mapping.

First, we map only worm interactions onto the yeast genome. We assess the predicting accuracy of our method, best-match and generalized interolog mapping against sets of gold standard positives ***P*** and negatives ***N*** (see Methods). Fig. 1A shows the relationship between accuracy and coverage in the worm-yeast mapping. The blue line indicates the accuracy of generalized interolog mapping from $J_E \leq 10^{-190}$ to 10^{-10}. The green line indicates the accuracy of best-match mapping at $J_E \leq 10^{-10}$. There are three clear observations:

1. While selecting only pairs of top ***R*** homologs (see Methods) as candidate interactions (i.e., rank-based interologs), the accuracy would be usually better than the accuracy of generalized interolog mapping. For example, the purple line consists of plots $Acc(10^{-70},\boldsymbol{R})$, $\boldsymbol{R} \in$ [1,5,10,15,20,25,30,35,40,45,50,55,60,65,70,75,80,85,90,

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

95, 100,'All'], $J_E \leq 10^{-70}$ was used as a good threshold of predicting interactions in Yu et al. [1]. The accuracy of $\boldsymbol{R}$ = 1, 5, and 10 are 0.22, 0.26, and 0.21, respectively, are better than $Acc(10^{-70}$, 'All') = 0.11. 'All' means $\boldsymbol{R}$ has no limit.

2. If $J_E \leq 10^{-110}$, $Acc(J_E$, 'All') will raise sharply but the number of true positives are ≤ 25. In other words, a very low coverage of yeast interactions.

3. The max accuracy of best-match mapping is 0.26 at $J_E \leq 10^{-60}$, but the max number of true positives is only 50 (at $J_E \leq 10^{-10}$). Similarly, a low coverage of yeast interactions.

To gather better statistics, we map inter-actions in worm, fly (*D. melanogaster*), mouse, and human (*H. Sapiens*) onto the yeast genome, assessing them against our gold standards. We perform a similar analysis in Fig. 1B. The number of true positives dotted in Fig. 1B is sum of true positives in worm-yeast, fly-yeast, mouse-yeast, and human-yeast mappings. The accuracy is calculated by sum of true and false positives in the four mapping processes.

In Fig. 1B, the comparison among rank-based interolog, best-match, and generalized interolog mapping is similar to that in Fig. 1A. The accuracy of $\boldsymbol{R}$ = 1, 5, and 10 are 0.21, 0.17, and 0.12, respectively, are better than $Acc(10^{-70}$, 'All') = 0.04. Overall, rank-based interolog mapping has better predicting accuracy than generalized interolog mapping at $10^{-110} < J_E \leq 10^{-10}$. Although best-match and generalized interolog (with $J_E \leq 10^{-110}$) mapping have higher accuracy than our method, their predictions suffer from low coverage of the yeast interactions.

B. Analysis of rank-based interologs

Functional similarity between homologous protein pairs

The homologs of a query protein selected at a certain cutoff may be different in subcellular compartment, biological process, or function from the query protein. For quantitatively assessing the reliable homologous pairs in rank-based, best-match and generalized interologs, we construct sets of $\boldsymbol{P'}$ and $\boldsymbol{N'}$ by the GO annotations (see Methods). GO consortium provides a standardized vocabulary, in which three structured ontologies have been proposed, which allow the description of cellular component (CC), biological process (BP), and molecular function (MF) [23]. These annotations particularly allow for assessing the functional similarities of genes or their products.

Based on Wu et al. [24], we calculate the functional similarities between query (in the four organisms) and target (in yeast) interactions by using GO annotations. However, not all of proteins have GO annotations. Table 1 shows the percentage of $TP'(10^{-10}$, 'All') and $FP'(10^{-10}$, 'All') with the terms of CC and BP ontologies in each organism. Here $TP'(10^{-10}$, 'All') = $TP(10^{-10}$, 'All') ∩ $\boldsymbol{P'}$ and $FP'(10^{-10}$, 'All') = $FP(10^{-10}$, 'All') ∩ $\boldsymbol{N'}$ (see Methods). For example, in worm-yeast mapping, the number of $TP'(10^{-10}$, 'All') is 412. There is 52.3% (412/788) of $TP(10^{-10}$, 'All') has CC and BP annotations.

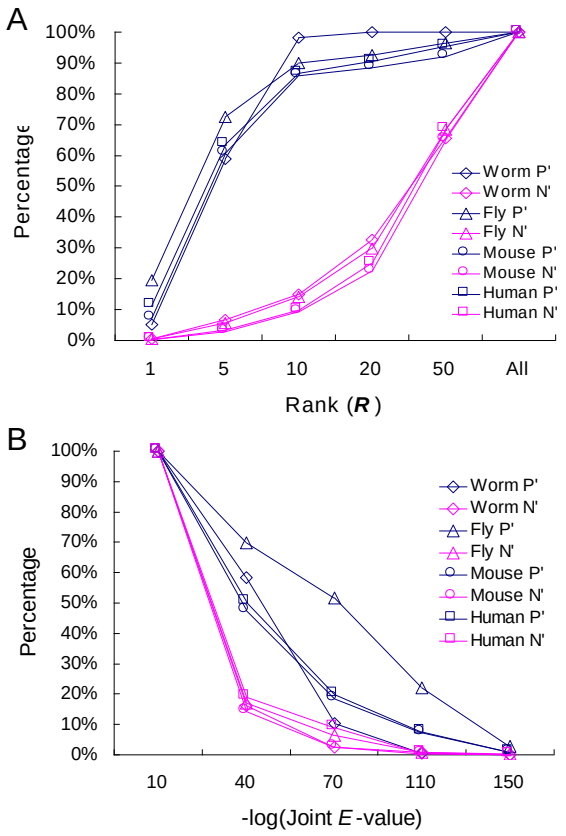

Fig. 2. The relationship between recall of TP'(10^{-10}, 'All') and FP'(10^{-10}, 'All') against (A) rank and (B) J_E. TP'(10^{-10}, 'All') and FP'(10^{-10}, 'All') of each mapping are represented with blue and pink lines, respectively.

The statistics of recalls of $TP'(10^{-10}$, 'All') and $FP'(10^{-10}$, 'All') in four mappings are showed in Fig. 2. Fig. 2A indicates the relationship between recall and rank. The recalls of $TP'(10^{-10}$, 'All') at $\boldsymbol{R}$ = 1, 5, 10 are 5.1% (21/412), 59.0% (243/412), and 98.3% (405/412), respectively. At $\boldsymbol{R}$ = 1, 5, 10, the recalls of $FP'(10^{-10}$, 'All') is 0.5% (15/2778), 6.7% (185/2778), and 14.8% (411/2778). There are similar trends in the four mappings from the source organisms to yeast.

Otherwise, there is no given JE could satisfy the demands together: High recall of true positives and low recall of false positives. For example, the recall of FP'(10^{-10}, 'All') is 16.3% at JE < 10^{-40}, near that at R = 10, but the recall of TP'(10-10, 'All') is only 58.3%. At JE < 10^{-40}, the recalls of true and false positives are 10.4% and 2.6%, respectively. These results suggest that rank-based mapping method could predict more reliable interactions under a given percentage of false positives than best-match and generalized interolog mapping methods.

TABLE 1. TRUE AND FALSE POSITIVES CASES SELECTED BY JE ON FOUR ORGANISMS

Species	$TP(10^{-10}$, 'All')	$FP(10^{-10}$, 'All')	$TP'(10^{-10}$, 'All')	$FP'(10^{-10}$, 'All')
Worm	788	13971	412 (52.3%)	2778 (19.9%)
Fly	780	73235	362 (46.4%)	23148 (31.6%)
Mouse	912	37636	685 (75.1%)	27770 (73.8%)
Human	2790	187149	1661 (59.5%)	128752 (68.8%)

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

Orthologous interactions

Fig. 3A and 3B shows the distribution of orthologous interactions against rank and J_E. Here "orthologous interaction" means a protein pair of orthologs and this pair is in ***P***. The total number of orthologous interactions of four mappings is 1626. Obviously, these orthologous interactions in four mappings concentrate in top-1–top-5 (totally 99.4%, 1616/1626; If top-1–top-10, 99.8%, 1622/1626) but spread in various J_E (e.g., 6.9% in $10^{-40} < J_E \leq 10^{-50}$ and 10.5% in $10^{-70} < J_E \leq 10^{-80}$). This results supply two suggestions: First, best-match mapping may be not good because it will lose ~44% of orthologous interactions. Second, generalized interolog mapping with any given J_E would lose part of orthologous interactions. Although loosing J_E could raise the coverage of orthologous interactions, the false positives would increase sharply. Our method could supply higher coverage of orthologous interactions and acceptable quantity of false positives.

In summary, we present two evidences for explaining why rank-based mapping could supply better predictions than other two methods. The first evidence is the assessment of homologous pairs evaluated by GO annotations. The second evidence is orthologous interactions between two organisms in four mappings. These results show that rank-based mapping method could predict more reliable interactions under a given percentage of false positives than best-match and generalized interolog mapping methods.

III. 3 Discussion

A. *Three types of good predictions*

We classify our predictions between organisms into three types to express why rank-based interolog mapping method could work. As $J_E \leq 10^{-70}$ was considered as a good threshold for predicting interactions, we represent the advantages of rank-based interologs in detail at $J_E \leq 10^{-70}$.

First, the true and false positives of a query interaction have various J_E. This suggests that any given J_E, such as 10^{-70}, may be not a good cutoff. For example, the query interaction P43686-P62195 in human has total 231 pairs (21 homologs × 21 homologs, excluding repeat pairs) of possible homologs (*E*-value $\leq 10^{-10}$) in yeast. Proteins P43686 and P62195 are 26S protease regulatory subunit 6B and subunit 8 of proteasome, respectively [25]. The total number of true positives (i.e., $|TP(10^{-10}, \text{'All'})|$) and false positives (i.e., $|FP(10^{-10}, \text{'All'})|$) are 22 and 23, respectively. In this case, the generalized interologs with $J_E \leq 10^{-70}$ have $|TP(10^{-70}, \text{'All'})| = 17$ and $|FP(10^{-70}, \text{'All'})| = 0$, the predicting accuracy is 1.0. For the rank-based interologs with $\boldsymbol{R} = 5$ and 10, $|TP(10^{-10}, 5)| = 15$ and $|FP(10^{-10}, 5)| = 0$, $|TP(10^{-10}, 10)| = 16$ and $FP(10^{-10}, 10) = 1$. The $Acc(10^{-10}, 5)$ and $Acc(10^{-10}, 10)$ are 1.0 and 0.94, lightly lower than $Acc(10^{-70}, \text{'All'})$. However, if considering the reliable interactions evaluated by GO annotations, the accuracies at $\boldsymbol{R} = 5$ and 10 are both better than $Acc(10^{-70}, \text{'All'})$ (see Fig. 4A).

In the second type, true positives of a query interaction have J_E from higher to lower than a given cutoff (e.g., 10^{-70}) and there are no or few false positives. Most of true positives are ranked in top ***R***. For this type, the query interaction O17071-Q09583 in worm is used as an example. O16368 is 26S protease regulatory subunit 4. Q9GZH5 is non-ATPase protein 1 of proteasome regulatory particle [26]. O17071 has 21 homologs (*E*-value $\leq 10^{-10}$) and Q09583 has 7 homologs (*E*-value $\leq 10^{-10}$) in yeast, respectively. $|TP(10^{-10}, \text{'All'})|$ is 49 in the total 147 homolog pairs. In this case, $|TP(10^{-70}, \text{'All'})| = 7$ and $|FP(10^{-70}, \text{'All'})| = 0$, the predicting accuracy of using threshold $J_E \leq 10^{-70}$ is 0.14. Otherwise, $|TP(10^{-40}, \text{'All'})| = 43$ and $|FP(10^{-40}, \text{'All'})| = 0$, $|TP(10^{-10}, 5)| = 25$ and $|FP(10^{-10}, 5)| = 0$, $|TP(10^{-10}, 10)| = 49$ and $|FP(10^{-10}, 10)| = 0$. The accuracy of $\boldsymbol{R} = 5$ and 10 are 0.51 and 1.0, respectively.

Third, all pairs of $TP(10^{-10}, \text{'All'})$ of a query interaction have higher J_E than a given cutoff, such as 10^{-70}. For this type, we get interaction O44156-Q27488 as an example. Proteins O44156 and Q27488 are proteasome subunit alpha 6 and subunit alpha 2, respectively [26]. Both O44156 and Q27488 have 7 possible homologs (*E*-value $\leq 10^{-10}$) in yeast. The minimum J_E of all pairs of homologs is 7.7×10^{-63}. In other words, there is no true positives has $J_E \leq 10^{-70}$, $|TP(10^{-70}, \text{'All'})| = 0$. Otherwise, $|TP(10^{-10}, 5)| = 15$ and $|FP(10^{-10}, 5)| = 0$, $|TP(10^{-10}, 10)| = 21$ and $|FP(10^{-10}, 10)| = 0$, the accuracy of rank-based interolog mapping method at $\boldsymbol{R} = 5$ and 10 are 0.71 and 1.0, respectively.

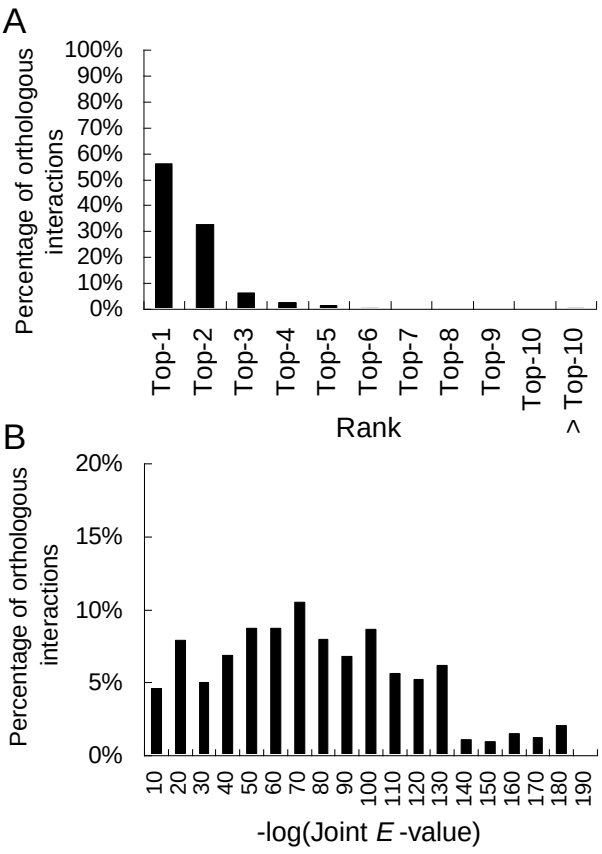

Fig. 3. Distributions of total orthologous inter-actions of four mappings against (A) rank and (B) J_E.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

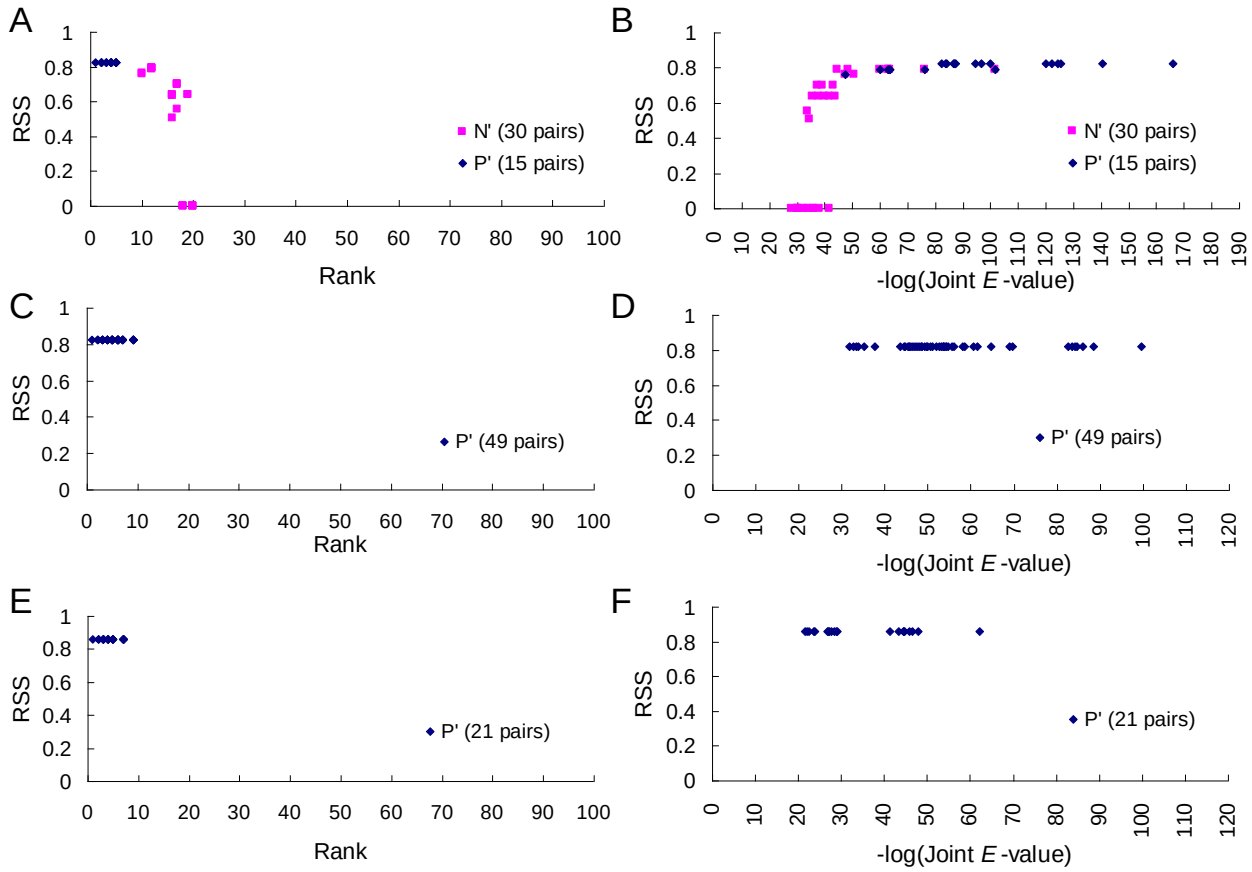

Fig. 4. Three cases of rank-based interolog mapping. (A) and (B) show the TP'(10^{-10}, 'All') (colored blue) and FP'(10^{-10}, 'All') (colored pink) of the query interaction P43686-P62195 in human. Similarly, (C) and (D) show the TP'(10^{-10}, 'All') and FP'(10^{-10}, 'All') of the query interaction O16368-Q9GZH5 in worm. (E) and (F) show the TP'(10^{-10}, 'All') and FP'(10^{-10}, 'All') of the query interaction O44156-Q27488 in worm. RSS is $RSS^{GO}_{A-B,A'-B'}$.

B. Case analyses

Furthermore, we analyze the three real cases in detail. In first case, Fig. 4A shows that all of 15 true positives ($0.8 < RSS^{GO}_{A-B,A'-B'} \leq 1.0$) are $\leq$ top 10. 97% (29/30) of false positives with $RSS^{GO}_{A-B,A'-B'} \leq 0.8$ are out of top 10. The rank of each pair is calculated in the same way described in Fig. 3. Comparing to Fig. 4B, reliable true positives spread in $10^{-170} \leq J_E \leq 10^{-40}$, this suggests that any given J_E would lose part of true positives.

Similarly, Fig. 4C and 4D represent the second case, interaction O16368-Q9GZH5 in worm. All of 49 true positives ($0.8 < RSS^{GO}_{A-B,A'-B'} \leq 1.0$) are $\leq$ top 10. These reliable true positives spread in $10^{-130} \leq J_E \leq 10^{-30}$. In this case, there are no true and false positives with $RSS^{GO}_{A-B,A'-B'} \leq 0.8$. Additionally, Fig. 4E and 4F show the third case O44156-Q27488 in worm. All of 21 true positives ($0.8 < RSS^{GO}_{A-B,A'-B'} \leq 1.0$) are $\leq$ top 10. These reliable true positives spread in $10^{-65} \leq J_E \leq 10^{-20}$. In this case, there are no true and false positives with $RSS^{GO}_{A-B,A'-B'} \leq 0.8$.

IV. CONCLUSIONS

In this paper, we propose a rank-based interolog mapping method for predicting interactions across species. Our method selects the pairs with higher sequence similarity instead of only best matches or all possible homologous pairs. Four mappings of worm-yeast, fly-yeast, mouse-yeast, and human-yeast are included in this study. Our results present that rank-based mapping method could predict more reliable interactions (including positives annotated by CC and BP ontologies and orthologous interactions) under a given percentage of false positives than best-match and generalized interolog mapping methods. In addition, based on above results, we suggest that ***R*** = 10 is a good threshold for predicting protein-protein interactions.

V. METHODS

A. Rank-based interolog mapping

Interolog mapping is a process that maps interactions in the source organism onto the target organism to predict possible interactions. To address the three questions of best-match and generalized interolog mapping described above, we introduce a new "rank-based interolog mapping" method using

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

part of possible homologs of interacting proteins. Operationally, homologs could be defined as the proteins having an *E*-value $\leq 10^{-10}$ from BLASTP [7, 27]. An overview of the rank-based interolog mapping is depicted in Fig. 5. The steps are described as following:

1. These possible homologs of proteins A and B in the target organism (yeast) are ranked by their *E*-values from low to high (i.e., from 0 to 10^{-10}), respectively.
2. These homologs ranked in top ***R*** are selected to pair with each other. These possible protein pairs between the homologs $A'_1 \ldots A'_R$ and $B'_1 \ldots B'_R$ are called ranked-based interologs. Otherwise, the all pairs between the homologs $A'_1 \ldots A'_m$ and $B'_1 \ldots B'_n$ are generalized interologs.
3. For any given protein in the source organism (e.g., worm), we collect all of its homologs by BLASTP *E*-value $\leq 10^{-10}$.

The best-match mapping method considers pairs between the best-matching homologs as the candidates of interactions [7]. The generalized interolog mapping method uses all pairs of homologs, which have joint similarities larger than a certain cutoff, to find possible interactions in the target organism [1]. In this study, we consider the protein pairs between the top ***R*** possible homologs as the candidates of interaction in the target organism.

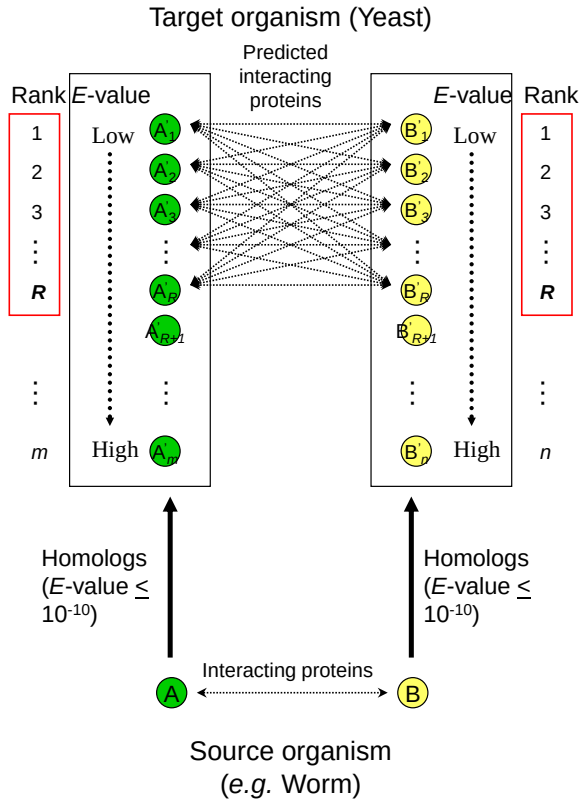

Fig. 5. Schematic illustration of rank-based interolog mapping method. Proteins A1', A2', …, Am' and B1', B2', …, Bn' are possible homologs (E-value < 10^{-10}) of proteins A and B in the source organism, respectively. All possible pairs between homologs A'1…A'R and B'1…B'R are called ranked-based interologs.

B. Source data sets

To assess the rank-based interolog mapping method, we need source organisms with known interaction data. In this study, *C. elegans*, *D. melanogaster*, *M. musculus*, and *H. Sapiens* are used as source organisms. We collect the interactions of these four organisms recorded in IntAct database (December 1, 2007) [16] (Table 2). We then map these interactions onto the yeast genome. The protein sequences of these four source organisms and the target organism yeast are from SWISS-PROT and SGD database [28], respectively.

C. Gold standard target data sets

Set of gold standard positives (P)

To assess the performance of interolog mapping, we need a collection of known interactions as positives in the target organism. Previously, a data set derived from the MIPS complex catalog, which contains 8,250 unique interacting protein pairs, has been used as a standard reference for known interactions [1, 4, 29]. We also consider the MIPS interactions as gold standard positives in this paper.

Table 2. Source data sets derived from IntAct

Species	Worm	Human	Fly	Mouse	Total
Number of interactions	4653	18943	19774	2728	46098

Set of gold standard positives (N)

A set of negatives (i.e., non-interacting proteins) in yeast is necessary for evaluating our method. Jansen et al. [5] considered pairs of proteins in different subcellular compartments as good estimates for non-interacting proteins. This set has 2,708,746 such protein pairs. Therefore, we find that 3,689 interactions in this set are also recorded in the core database of DIP [15]. We exclude these interactions and take 2,705,057 protein pairs as the set of gold standard negatives in this study.

D. Accuracy of interolog mapping

We assess the predicting accuracy of our method, best-match and generalized interolog mapping against ***P*** and ***N*** in yeast. The accuracy (*Acc*) is calculated as following:

$$Acc(J,R) = \frac{TP(J,R)}{TP(J,R) + FP(J,R)}$$

In this equation, $TP(J,R) = H(J,R) \cap \boldsymbol{P}$, $FP(J,R) = H(J,R) \cap \boldsymbol{N}$. $H(J,R)$ means the sets of rank-based interologs, best-matching homologous pairs, or generalized interologs in yeast at a certain cutoff. For example, in rank-based interolog mapping, J is a given joint *E*-value (see below) and R is the number of homologs selected by ranking (i.e., top ***R***). Otherwise, in generalized interolog mapping, J is a certain joint *E*-value and ***R*** has no limits. $|TP(J,R)|$ and $|FP(J,R)|$ are the number of true and false positives at a given J and ***R***.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

E. Joint E-value (J_E)

J_E is the geometric means of *E*-values for the two pairs of interacting proteins. For example, if the *E*-values of A-A' and B-B' are $E_{A\text{-}A'}$ and $E_{B\text{-}B'}$, J_E between pairs A-B and A'-B' is $J_E = \sqrt{E_{A-A'} \times E_{B-B'}}$.

F. GO similarity measure

We assume that if a pair A'-B' in yeast is a reliable homologous pairs of A-B, A-A' and B- B' would be in similar subcellular compartment, biological process and have similar molecular function. Wu et al. [24] proposed a method to measure semantic similarity between two proteins by using CC and BP annotations. Based on the relative specificity similarities (RSS) defined by their method, we calculate the similarity in cellular component and biological process between a protein pair A-B and its rank-based interologs A'-B' ($RSS^{GO}_{A-B,A'-B'}$). RSS values for the CC (RSS^{CC}) and BP (RSS^{BP}) ontologies mean the similarity of CC and BP terms of a given protein pair, respectively. The values are between 0 and 1.0. The equations we used are as follows.

$$RSS^{CC}_{A-B,A'-B'} = \sqrt{RSS^{CC}_{A-A'} \times RSS^{CC}_{B-B'}}$$

$$RSS^{BP}_{A-B,A'-B'} = \sqrt{RSS^{BP}_{A-A'} \times RSS^{BP}_{B-B'}}$$

$$RSS^{GO}_{A-B,A'-B'} = \sqrt{RSS^{CC}_{A-B,A'-B'} \times RSS^{BP}_{A-B,A'-B'}}$$

Wu et al. [24] supplied both three confidence levels of yeast protein pairs annotated in the CC and BP ontologies. Their results showed that 78% interactions of their positive dataset fall into the high-confidence segment of $0.8 < RSS^{CC} \leq 1.0$ and $0.8 < RSS^{BP} \leq 1.0$. They suggested that the highest-confidence segment might contain most yeast protein-protein interactions. We use the thresholds to construct two datasets, ***P'*** and ***N'***. ***P'*** is the interaction dataset including these protein-protein interactions with $0.8 < RSS^{GO}_{A-B,A'-B'} \leq 1.0$ in ***P***. ***N'*** is the dataset consisted of ***N*** and these interactions of ***P*** but having $RSS^{GO}_{A-B,A'-B'} \leq 0.8$.

G. Orthologous interactions between source and target organisms

We identify the orthologous proteins between the source organisms and yeast by ENSEMBL database (Mar, 2008) [30]. For comparing the coverage of orthologous interactions (i.e., the protein pairs of orthologs and these pairs are in ***P***) of our method, best-match and generalized interolog mapping, we identify and count the interacting pairs of orthologs in all pairs of possible homologs (*E*-value $\leq 10^{-10}$). For example, interacting proteins P47857-P12382 in mouse have an orthologous interaction YMR205C-YGR240C. In Fig. 3A, the label "Top-*n*" is the maximum of ranks of $E_{A\text{-}A'}$ and $E_{B\text{-}B'}$. For example, YMR205C and YGR240C are the top 1 and top 2 in the ranking of homologs by *E*-values, respectively. The label of pair YMR205C-YGR240C is $n = \max(1, 2) = 2$.

REFERENCES

[1] H. Yu, N. Luscombe, H. Lu, X. Zhu, Y. Xia, J. Han, *et al.*, "Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs," *Genome Res.*, vol. 14, pp. 1107-1118, 2004.

[2] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proceedings of the National Academy of Science of the USA*, vol. 98, pp. 4569-4574, 2001.

[3] A. Pandey and M. Mann, "Proteomics to study genes and genomes," *Nature*, vol. 405, pp. 837-846, 2000.

[4] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, *et al.*, "Comparative assessment of large-scale data sets of protein-protein interactions," *Nature*, vol. 417, pp. 399-403, 2002.

[5] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, *et al.*, "A Bayesian networks approach for predicting protein-protein interactions from genomic data," *Science*, vol. 302, pp. 449-453, 2003.

[6] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates, "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles," *Proceedings of the National Academy of Science of the USA*, vol. 96, pp. 4285-4288, 1999.

[7] L. R. Matthews, P. Vaglio, J. Reboul, H. Ge, B. P. Davis, J. Garrels, *et al.*, "Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs"," *Genome Research*, vol. 11, pp. 2120-2126, 2001.

[8] J. Wojcik and V. Schachter, "Protein-protein interaction map inference using interacting domain profile pairs," *Bioinformatics*, vol. 17, pp. S296-S305, 2001.

[9] P. Aloy and R. B. Russell, "Interrogating protein interaction networks through structural biology," *Proceedings of the National Academy of Science of the USA*, vol. 99, pp. 5896-5901, 2002.

[10] M. P. Cary, G. D. Bader, and C. Sander, "Pathway information for systems biology," *FEBS Letters*, vol. 579, pp. 1815-1820, 2005.

[11] Y.-C. Chen, Y.-S. Lo, W.-C. Hsu, and J.-M. Yang, "3D-partner: a web server to infer interacting partners and binding models," *Nucleic Acids Research*, pp. W561-W567, 2007 2007.

[12] J. Sun, Y. Sun, G. Ding, Q. Liu, C. Wang, Y. He, *et al.*, "InPrePPI: an integrated evaluation method based on genomic context for predicting protein-protein interactions in prokaryotic genomes," *BMC Bioinformatics*, vol. 8, p. 414, 2008.

[13] C. von Mering, L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini, *et al.*, "STRING: known and predicted protein-protein associations, integrated and transferred across organisms," *Nucleic Acids Research*, vol. 33, pp. D433-D437, 2005.

[14] H. Mewes, D. Frishman, U. Güldener, G. Mannhaupt, K. Mayer, M. Mokrejs, *et al.*, "MIPS: a database for genomes and protein sequences," *Nucleic Acids Res.*, vol. 30, pp. 31-34, 2002.

[15] I. Xenarios, L. Salwiński, X. Duan, P. Higney, S. Kim, and D. Eisenberg, "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions," *Nucleic Acids Res.*, vol. 30, pp. 303-305, 2002.

[16] S. Kerrien, Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, *et al.*, "IntAct-open source resource for molecular interaction data," *Nucleic Acids Research*, vol. 35, pp. D561-D565, 2007.

[17] B. Kelley, R. Sharan, R. Karp, T. Sittler, D. Root, S. BR, *et al.*, "Conserved pathways within bacteria and yeast as revealed by global protein network alignment," *Proc. Natl. Acad. Sci. U. S. A.* , vol. 100, pp. 11394-11399, 2003.

[18] R. Sharan, S. Suthram, R. Kelley, T. Kuhn, S. McCuine, P. Uetz, *et al.*, "Conserved patterns of protein interaction in multiple species," *Proc. Natl. Acad. Sci. U. S. A.* , vol. 102, pp. 1974-1979, 2005.

[19] L. Lu, H. Lu, and J. Skolnick, "MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading," *Proteins: Structure, Function and Bioinformatics*, vol. 49, pp. 350-364, 2002.

[20] M. Egerton, O. Ashe, D. Chen, B. Druker, W. Burgess, and L. Samelson, "VCP, the mammalian homolog of cdc48, is tyrosine phosphorylated in response to T cell antigen receptor activation," *EMBO J.*, vol. 11, pp. 3533-3540, 1992.

[21] O. Gadal, D. Strauss, J. Braspenning, D. Hoepfner, E. Petfalski, P. Philippsen, *et al.*, "A nuclear AAA-type ATPase (Rix7p) is required for

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

biogenesis and nuclear export of 60S ribosomal subunits," *EMBO J.*, vol. 20, pp. 3695-3704, 2001.
[22] I. Tirosh and N. Barkai, "Computational verification of protein-protein interactions by orthologous co-expression," *BMC Bioinformatics*, vol. 6, p. 40, 2005.
[23] M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, *et al.*, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat. Genet.*, vol. 25, pp. 25-29, 2000.
[24] X. Wu, L. Zhu, J. Guo, D. Zhang, and K. Lin, "Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations," *Nucleic Acids Res.*, vol. 34, pp. 2137-2150, 2006.
[25] A. Davy, P. Bello, N. Thierry-Mieg, P. Vaglio, J. Hitti, L. Doucette-Stamm, *et al.*, "A protein-protein interaction map of the Caenorhabditis elegans 26S proteasome," *EMBO Rep.*, vol. 2, pp. 821-828, 2001.
[26] C. e. S. Consortium, "Genome sequence of the nematode C. elegans: a platform for investigating biology," *Science*, vol. 282, pp. 2012-2018, 1998.
[27] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic local alignment search tool," *J Mol. Biol.*, vol. 215, pp. 403-410, 1990.
[28] S. Weng, Q. Dong, R. Balakrishnan, K. Christie, M. Costanzo, K. Dolinski, *et al.*, "Saccharomyces Genome Database (SGD) provides biochemical and structural information for budding yeast proteins," *Nucleic Acids Res.*, vol. 31, pp. 216-218, 2003.
[29] A. Edwards, B. Kus, R. Jansen, D. Greenbaum, J. Greenblatt, and M. Gerstein, "Bridging structural biology and genomics: assessing protein interaction data with known complexes," *Trends Genet.*, vol. 18, pp. 529-536, 2002.
[30] E. Birney, T. Andrews, P. Bevan, M. Caccamo, Y. Chen, L. Clarke, *et al.*, "An Overview of Ensembl," *Genome Res.*, vol. 14, pp. 925-928, 2004.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE