Enhancing the accuracy of HMM-based conserved pathway prediction using global correspondence scores

Xiaoning Qian*[1], Sayed Mohammad Ebrahim Sahraeian[2] , Byung-Jun Yoon*[2]

[1]Department of Computer Science and Engineering, University of South Florida, Tampa, FL, 33620, USA
[2]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, 77843, USA

Email: XQ*: xqian@cse.usf.edu; SMES : msahraeian@tamu.edu; BJY*: bjyoon@ece.tamu.edu;

*Corresponding author

Abstract

Background: Comparative network analysis aims to identify common subnetworks in biological networks. It can facilitate the prediction of conserved functional modules across different species and provide deep insights into their underlying regulatory mechanisms. Recently, it has been shown that hidden Markov models (HMMs) can provide a flexible and computationally efficient framework for modeling and comparing biological networks.

Results: In this work, we show that using global correspondence scores between molecules can improve the accuracy of the HMM-based network alignment results. The global correspondence scores are computed by performing a semi-Markov random walk on the networks to be compared. The resulting score naturally integrates the sequence similarity between molecules and the topological similarity between their molecular interactions, thereby providing a more effective measure for estimating the functional similarity between molecules. By incorporating the global correspondence scores, instead of relying on sequence similarity or functional annotation scores used by previous approaches, our HMM-based network alignment method can identify conserved subnetworks that are functionally more coherent.

Conclusions: Performance analysis based on synthetic and microbial networks demonstrates that the proposed network alignment strategy significantly improves the robustness and specificity of the predicted alignment results, in terms of conserved functional similarity measured based on KEGG ortholog (KO) groups. These results clearly show that the HMM-based network alignment framework using global correspondence scores can effectively find conserved biological pathways and has the potential to be used for automatic functional annotation of biomolecules.

Background

With the increasingly high coverage of molecular interactions owing to the advancement of high-throughput techniques for measuring biomolecular interactions, such as the two-hybrid screening [1] and co-immunoprecipitation [2], comparative analysis of biological networks has recently attracted significant research attention. It has been demonstrated that comparative network analysis can provide an effective means of systematically studying molecular interactions in various organisms and gaining novel system-level insights [3–18]. For example, local network alignment across different species can identify similar subnetwork regions in the respective networks, which may lead to the discovery of conserved pathways that carry out essential cellular functionalities [3, 5, 6, 9, 11, 15, 16, 19]. The concept of comparative network analysis can lead to the development of novel computational tools that allow us to transfer biological knowledge across species, especially from well-studied species to less-studied species [19].

Current local network algorithms [3, 5, 6, 9, 15] search for similar subnetwork regions by optimizing a pre-defined alignment score that incorporates the *topological similarity* of the interaction patterns in the compared networks as well as the *node similarity* of the molecules that belong to different networks, typically measured based on sequence similarity. To obtain better alignment results that are biologically more significant, there have been research efforts to improve the scoring scheme by incorporating evolutionary [4] or functional relationships [11, 16] between molecules. Although there are various approaches for measuring the similarity between network nodes, most of the existing approaches compute this similarity based on the properties of individual nodes, such as their composition, functionality, or evolutionary relationships. However, cellular functions are carried out by collaborative efforts among many molecules, where interacting molecules may carry similar functionalities and share common characteristics. Therefore it would be reasonable to expect that, when evaluating the node similarity, incorporating additional information about the interacting molecules would enhance the network alignment results and lead to predictions that are biologically more meaningful.

Recently, we have introduced an effective framework for local network alignment based on hidden Markov models (HMMs), in which we integrate both the node sequence similarity and the interaction reliability into the scoring scheme by determining the parameters of the HMMs correspondingly [15]. We also developed an efficient dynamic programming algorithm that can find the closest pair of pathways from the respective networks in polynomial time. The HMM-based local alignment method can deal with a large

class of path isomorphism and it allows one to search for long conserved pathways across large-scale networks. In this paper, we implement a semi-Markov random walk framework that diffuses the relationships of all the molecule pairs across the networks to obtain a *global correspondence score* between every pair of nodes. The resulting global correspondence score reflects the global similarity between nodes in different networks, by seamlessly integrating the topological similarity and individual node similarity. Alignment results based on synthetic networks and microbial protein-protein interaction (PPI) networks show that the performance of the HMM-based local alignment scheme can be significantly improved by utilizing the *global* correspondence score instead of the original *individual* sequence similarity score. The major contributions of this paper include the following: first, we integrate the global node correspondence scoring scheme into the HMM-based local network alignment framework [15], which leads to more accurate and robust alignment results; second, we thoroughly evaluate the performance of the proposed scheme based on synthetic benchmark networks, as well as real microbial networks, which clearly demonstrates the advantages of utilizing global correspondence scores, especially, in combination with the HMM-based framework.

Methods

Local network alignment based on hidden Markov models

In this section, we briefly review our local network alignment algorithm based on hidden Markov models (HMMs) [14, 15]. We focus on aligning two biological networks to identify the common pathways that are conserved in both networks. Suppose we have two biological networks, represented as two graphs $\mathcal{G}_1 = (\mathcal{U}, \mathcal{D})$ and $\mathcal{G}_2 = (\mathcal{V}, \mathcal{E})$. In graph $\mathcal{G}_1$, $\mathcal{U} = \{u_1, u_2, \cdots, u_{N_1}\}$ of N_1 nodes represents the corresponding molecules, and $\mathcal{D} = \{d_{ij}\}$ of M_1 edges indicates the presence of interactions d_{ij} between the two molecules u_i and u_j. Similarly, we assume that $\mathcal{G}_2$ has a set $\mathcal{V} = \{v_1, v_2, \cdots, v_{N_2}\}$ of N_2 nodes and a set $\mathcal{E} = \{e_{ij}\}$ of M_2 edges. We denote the interaction reliability score between u_i and u_j in $\mathcal{G}_1$ as $w_1(u_i, u_j)$ and the interaction reliability between v_i and v_j in $\mathcal{G}_2$ as $w_2(v_i, v_j)$. The node similarity between $u_i \in \mathcal{U}$ and $v_j \in \mathcal{V}$ is denoted as $s(u_i, v_j)$.

In order to use HMMs to search for the pathways that are conserved in both networks, we search for the best matching pair of paths $\mathbf{u} = u_1 u_2 \dots u_L$ $(u_i \in \mathcal{U})$ and $\mathbf{v} = v_1 v_2 \dots v_L$ $(v_j \in \mathcal{V})$ of length L in the respective networks that maximizes the pathway alignment score $H(\mathbf{u}, \mathbf{v})$. The alignment score $H(\mathbf{u}, \mathbf{v})$ integrates the *node similarity score* $s(u_i, v_j)$ between the aligned nodes u_i and v_j $(1 \le i, j \le L)$, the

interaction reliability score $w_1(u_i, u_{i+1})$ between u_i and u_{i+1} $(1 \leq i \leq L-1)$, the interaction reliability score $w_2(v_j, v_{j+1})$ between v_j and v_{j+1} $(1 \leq j \leq L-1)$, and the penalty for potential gaps in the alignment.

We first construct two HMMs respectively for two given networks. For $\mathcal{G}_1$, we design the state transition diagram of its corresponding HMM based on the graph structure of $\mathcal{G}_1$. The resulting HMM contains a hidden state for each node $u_i \in \mathcal{U}$, which we also denote as u_i for convenience. State transition is allowed from u_i to u_j for (u_i, u_j) such that $d_{ij} \in \mathcal{D}$. The HMM for $\mathcal{G}_2$ can be constructed in a similar way. To allow flexible node insertions and/or deletions in the alignment result, we add auxiliary states to the HMMs as described in [14, 15]. The state transition probabilities of the HMMs are determined based on the interaction reliability scores $w_1(u_i, u_{i+1})$ and $w_2(v_j, v_{j+1})$. By introducing a "virtual observation sequence" $\mathbf{q} = q_1 \cdots q_L$ that is *jointly* emitted by the two HMMs, we design the emission probabilities based on the node similarity $s(u_i, v_j)$. Using these HMMs, the problem of finding the optimal pair of paths in the two networks is translated into that of finding the optimal pair of state sequences in the two HMMs that jointly maximize the probability $P(\mathbf{q}, \mathbf{u}, \mathbf{v})$ of the "virtual observation sequence":

$$(\mathbf{u}^*, \mathbf{v}^*) = \arg\max_{(\mathbf{u},\mathbf{v})} P(\mathbf{q}, \mathbf{u}, \mathbf{v}).$$

We can use $\log P(\mathbf{q}, \mathbf{u}, \mathbf{v})$ as the alignment score $H(\mathbf{u}, \mathbf{v})$, and find the best matching pair of paths using dynamic programming [15]. For this purpose, we first define the score for the most probable pair of paths of length $t\,(\leq L)$ as follows:

$$\gamma(t, j, \ell) = \max_{i,k} \Big[\gamma(t-1, i, k) + t_{w_1}(u_i, u_j) + t_{w_2}(v_k, v_\ell) + s(u_j, v_\ell)\Big], \tag{1}$$

where $t_{w_1}(u_i, u_j)$ and $t_{w_2}(v_k, v_\ell)$ are the logarithms of transition probabilities determined by the interaction reliability scores in the respective networks. Next, we find the optimal pair of paths $(\mathbf{u}^*, \mathbf{v}^*)$

$$H(\mathbf{u}^*, \mathbf{v}^*) \quad = \quad \max_{\mathbf{u},\mathbf{v}} \Big[H(\mathbf{u}, \mathbf{v})\Big] = \max_{j,\ell} \gamma(L, j, \ell), \tag{2}$$

by iteratively computing the score in (1) for $\ell = 1, 2, \cdots, L$. Instead of finding only the best matching pair of paths, we can also search for the top k path pairs by replacing the max operator in (2) by an operator that finds the k largest scores. The computational complexity of the described dynamic programming algorithm is only $O(kLM_1M_2)$ for finding the top k pairs of matching paths. Note that the computational complexity is linear with respect to each parameter k, L, M_1, and M_2.

In our previous implementation of HMM-based local alignment [14, 15, 20], we have used the sequence similarity between individual molecules to measure the node similarity $s(u_i, v_j)$. As we discussed earlier, it

is desirable to integrate all the available information to measure the similarity between network nodes, instead of relying on the similarity between individual molecules. In this paper, we propose to use a semi-Markov random walk model to define a global correspondence scoring scheme for measuring node similarity by incorporating the topological properties around the nodes. As we will demonstrate later, the use of global correspondence scores can improve the accuracy and robustness of the HMM-based alignment results.

Computation of global correspondence scores through semi-Markov random walk

In order to predict the global correspondence between nodes, we should first consider the similarity between the corresponding molecules themselves, in terms of sequence, structure, and/or function. However, considering that biomolecules carry out their functions through intertwined interactions with other molecules, it is important to consider these interaction patterns as well when evaluating the global similarity between nodes. As recently proposed and discussed in [10, 18, 21, 22], Markov random walk can provide an elegant framework for evaluating the global correspondence between nodes that belong to different networks by seamlessly integrating the similarity between the nodes themselves and that between their interaction patterns.

In this work, we adopt the semi-Markov random walk approach [18] to compute the global correspondence scores for the node similarity $s(u_i, v_j)$. The basic idea of this scheme is to perform a simultaneous semi-Markov random walk on $\mathcal{G}_1$ and $\mathcal{G}_2$, such that the random walker moves to one of the neighboring nodes in each network at each time point. The next node is randomly selected among all the neighboring nodes, where nodes with higher interaction reliability have a larger chance to be selected. The time that the random walker spends at a given pair of nodes $u_i \in \mathcal{G}_1$ and $v_j \in \mathcal{G}_2$ is proportional to the sequence similarity between the nodes. According to this model, the long-run proportion of time that the random walker spends at (u_i, v_j) will increase if u_i and v_j have higher individual node similarity (e.g., sequence similarity). Furthermore, the proportion of time spent at (u_i, v_j) will also increase if the two nodes are surrounded by similar nodes, hence have a higher topological similarity. As a result, this semi-Markov random walk provides an elegant way of evaluating the global similarity between nodes by integrating individual node similarity and topological similarity. Using this model, we can compute the global correspondence score as follows:

$$s(u_i, v_j) = \frac{\pi_1(u_i)\pi_2(v_j)h(u_i, v_j)}{\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\pi_1(u_i)\pi_2(v_j)h(u_i, v_j)}, \tag{3}$$

in which $\pi_1(u_i)$ is the stationary probability of visiting node u_i in an ordinary Markov random walk on $\mathcal{G}_1$, $\pi_2(v_j)$ is the stationary probability of visiting v_j in a Markov random walk on $\mathcal{G}_2$, and $h(u_i, v_j)$ estimates the individual node similarity between u_i and v_j, which is measured in terms of sequence similarity in this work. The above scheme is conceptually similar to the one proposed in [10], where the similarity between two nodes in different networks are measured by *linearly* combining the topological similarity score and the sequence similarity score. The resulting score can be viewed as the long-run proportion of time spent at the given pair of nodes based on a "Markov random walk with restart" model, in which the restart probability has to be chosen in advance to balance the contributions from the interaction similarity and the sequence similarity, typically in an ad-hoc manner. Note that such parameter tuning is not needed in the semi-Markov random walk approach adopted in this work.

In the following sections, we analyze the effect of using the global correspondence scores in the HMM-based local network alignment method. More specifically, we evaluate the performance of the HMM-based local network alignment method when using the global correspondence score for $s(u_i, v_j)$ given in (3), and compare it to the performance of the HMM-based alignment method that directly uses the sequence similarity score with $s(u_i, v_j) = h(u_i, v_j)$, as originally proposed in [14, 15].

Results and Discussion

Aligning synthetic networks

To illustrate the advantages of using the global correspondence scores in the HMM-based local network alignment scheme, we first conducted a set of experiments based on synthetically generated networks. Figure 1(A) shows two simple synthetic networks, each of which contains a similar core path, respectively marked in dark blue and dark red. We added more nodes around the core path in each network in a way that the corresponding nodes in the core paths have similar local topological structures. We further assigned individual node similarity scores, where nodes in the respective networks that are located at similar vertical levels were given higher scores. We assigned exceptionally high similarity scores to two node pairs (u_4, v_{11}) and (u_8, v_{13}), which are shown by dashed lines in Fig. 1(A). These two highly similar pairs are analogous to molecules in real biological networks that have high sequence similarity without real biological significance. Such nodes can mislead the alignment algorithm, yielding inaccurate alignment results. In fact, we may typically face similar problems when aligning two or more large-scale biological networks. The network adjacency matrices and the node similarity matrix are provided in the

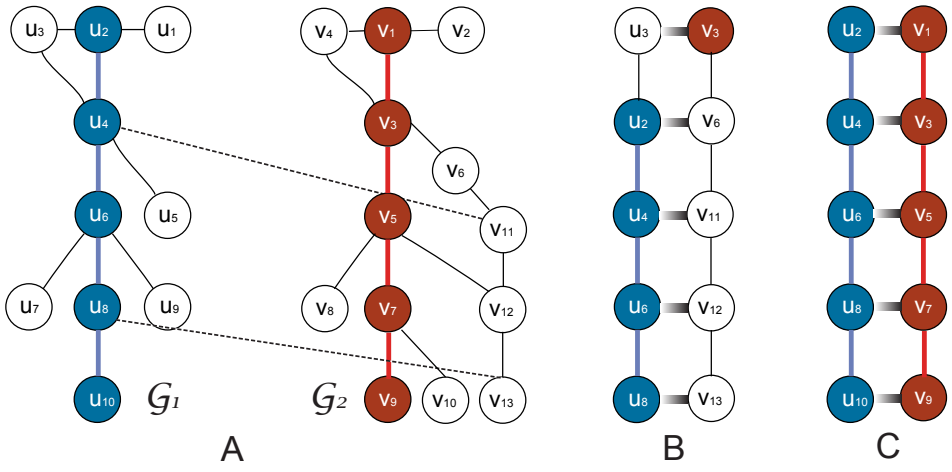

Figure 1: An illustrative example that demonstrates the advantage of using global correspondence scores: (A) Two small synthetic networks that contain similar paths (shown in colors); (B) The top pair of aligned paths predicted by the HMM-based alignment algorithm using the individual node similarity scores; (C) The top pair of aligned paths predicted by the HMM-based method using global node correspondence scores computed by semi-Markov random walk.

Supplementary materials.

We applied the HMM-based local alignment to identify the most similar pair of paths of length $L = 5$. The identified top pair of paths when directly using the assigned node similarity scores is shown in Fig. 1(B). We notice that the alignment result is strongly influenced by the high similarity pairs (u_4, v_{11}) and (u_8, v_{13}) in this case and the prediction does not capture the obvious topological similarity in the two networks. Next, we computed the global correspondence scores between nodes based on the semi-Markov random walk scheme and used these scores in the alignment algorithm, instead of the original node similarity scores. Figure 1(C) shows the top path alignment for this case, where the core paths were accurately identified as we expect based on the topology of the two networks. Simulations based on other small synthetic networks, constructed in similar ways, yielded similar results (see *Supplementary materials* for other examples) .

For a more thorough performance comparison between the two different schemes—the original scheme that directly uses the individual similarity scores and the proposed scheme that uses the global correspondence scores computed by semi-Markov random walk—we further created a benchmark set that consists of large synthetic networks generated based on a scale-free model [23]. Although we can also evaluate the performance of network alignment algorithms by aligning real biological networks and measuring the

accuracy of the alignment results using functional annotations based on Gene Ontology (GO) terms [24] or KEGG ortholog (KO) group annotations [25], these annotations are still highly incomplete and may not accurately reflect the real functional similarity between molecules. As a result, a carefully constructed synthetic benchmark dataset may provide a better benchmark for evaluating future network alignment algorithms.

To construct the synthetic networks, we first randomly generated an undirected seed network $\mathcal{G}$ of size 20 with an average degree of 10. Next, we grew this network according to the BA (Barabasi and Albert) model [23] to generate a random scale-free network using the preferential attachment algorithm [26]. In this algorithm, at each time step, a new node is added to the network and connected to m existing nodes with a probability that is proportional to the number of links that the nodes already have. As shown in [23], the resulting network captures several important characteristics of real PPI networks. The scale-free degree distribution is one such property, which means that the degree distribution of the network approximately follows the power law $P(k) \sim k^{\gamma}$, where γ is the degree exponent. In this work, we used this model with $m = 10$ to grow $\mathcal{G}$ to a network of size 1000. Once $\mathcal{G}$ was created, we duplicated the network into two identical networks $\mathcal{G}_1 = \mathcal{G}$ and $\mathcal{G}_2 = \mathcal{G}$. To model the functional coherence between orthologous proteins, we then assigned a distinct group annotation to each pair of corresponding proteins in the two networks. More specifically, both the node u_i in $\mathcal{G}_1$ and the node v_i in $\mathcal{G}_2$ were assigned to the ith functional group. We randomly assigned individual node similarity scores between orthologous nodes according to the Gaussian distribution $\mathcal{N}(\mu_o, \sigma_o^2)$ with mean $\mu_o = 300$ and standard deviation $\sigma_o = 100$. The node similarity scores between non-orthologous nodes were randomly assigned according to a different distribution $\mathcal{N}(\mu, \sigma^2)$, where $\sigma = 100$, and μ was used as a free parameter that determines the level of overlap between the two similarity score distributions. Node similarity scores below a certain threshold (set to 50 in this work) were set to zero. For every node, we also restricted the number of non-orthologous nodes with a nonzero similarity score to 10. These settings were used to make the resulting random networks similar to real PPI networks in public databases.

Up to this point, the two networks $\mathcal{G}_1$ and $\mathcal{G}_2$ were topologically identical. To introduce topological differences in these networks, we randomly deleted 10% of the edges in $\mathcal{G}_1$ and $\mathcal{G}_2$. Furthermore, we also randomly deleted 10% of the nodes in the two networks and added back an identical number of new nodes by growing the networks using the preferential attachment algorithm. No functional group was assigned to

these randomly inserted nodes. The node similarity between the inserted nodes in one network and the nodes in the other network was sparsely assigned according to $\mathcal{N}(\mu, \sigma^2)$, as before.

Based on the above model, we generated two networks $\mathcal{G}_1$ and $\mathcal{G}_2$ with $\mu = 200$. Using the HMM-based local network alignment algorithm, we identified the top 200 high-scoring path alignments with gaps. To find the top 200 path alignments, we iterated the following steps: (i) find the optimal path alignment; (ii) store the predicted alignment; (iii) remove the interactions included in the path alignments; (iv) repeat the experiment to find the next path alignment. The accuracy of the identified path pairs are measured based on the group annotations of the aligned nodes as in [11, 15]. We define the cumulative specificity of the top k alignments as follows:

$$cs_k = \frac{\sum_{i=1}^{k} c_i^c}{\sum_{i=1}^{k} c_i^a}, \tag{4}$$

where c_i^c the total number of correctly aligned node pairs in the top ith alignment, and c_i^a is the total number of annotated node pairs also in the top ith alignment. We also define the cumulative coverage of the top k alignments as $cc_k = \sum_{i=1}^{k} c_i^c$, which computes the cumulative number of pairs with the same functional annotations. This metric measures the size of the accurately aligned network regions covered by the top k path alignments. We repeated the alignment experiments for different path lengths: $L = 10, 20,$ and 30. The alignment results are summarized in Fig. 2. As we can see in this figure, the use of global node correspondence scores computed by the semi-Markov random walk approach significantly improves the specificity of the alignment, while the coverage is also slightly improved. This clearly shows that incorporating topological information of network nodes into the alignment process can be crucial in obtaining accurate alignment results when the individual node similarity scores are not highly discriminative by themselves.

Next, we investigated performance improvement by the proposed approach for varying levels of overlap between the node similarity score distribution for orthologous nodes and that for non-orthologous nodes. We set $L = 30$ and evaluated the performance of the HMM-based alignment algorithm using the global correspondence scores and the individual similarity scores, respectively, for various values of μ. The simulation results are shown Fig. 3. As shown in this figure, the advantage of using the global correspondence scores for network alignment becomes more prominent for larger μ. When μ is small (e.g., $\mu = 150$), we can accurately identify orthologous nodes by using individual node similarity scores. However, as μ increases and the individual node similarity scores become less discriminative, it becomes

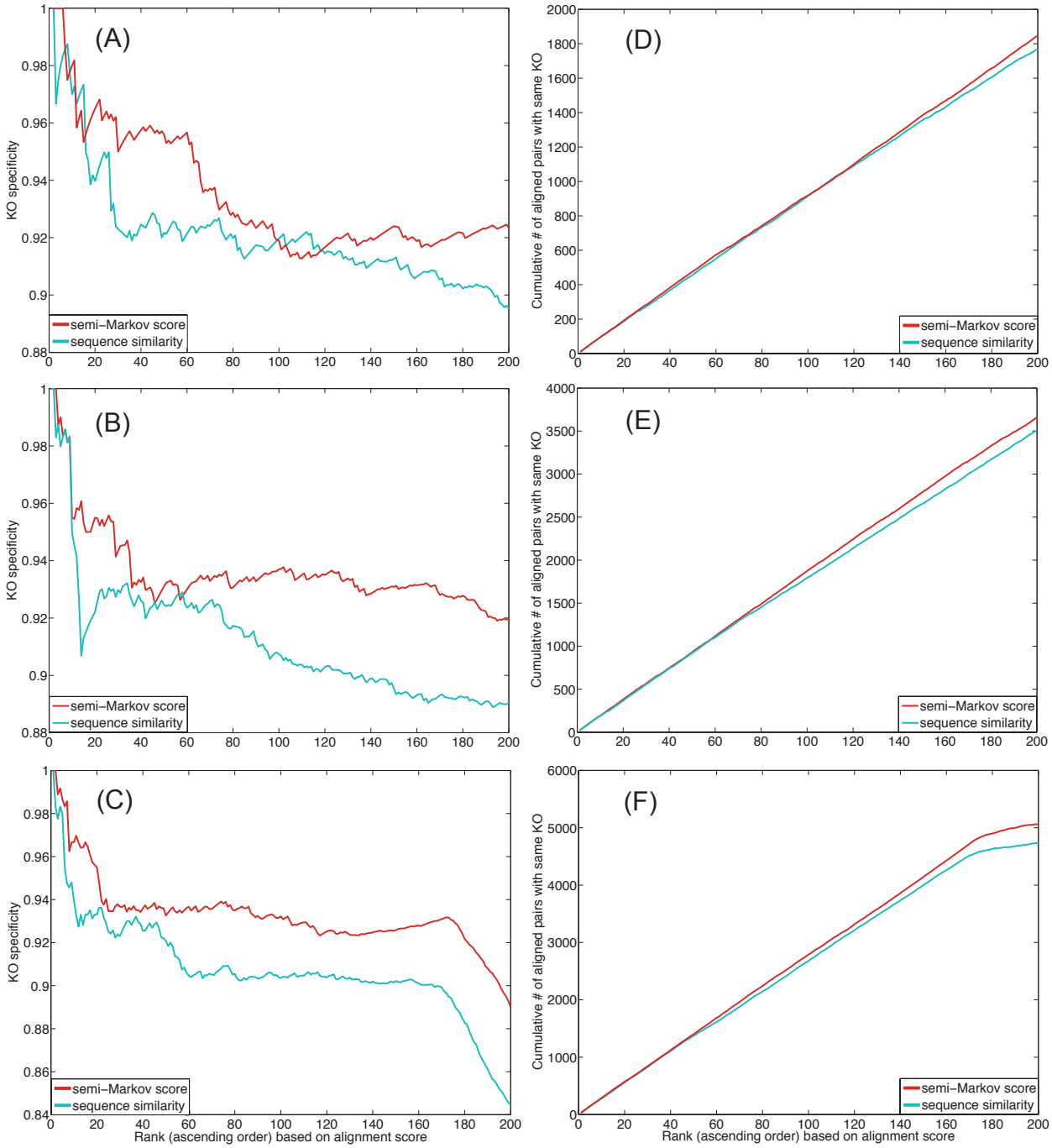

Figure 2: Functional specificity (A,B,C) and coverage (D,E,F) of the synthetic network alignment obtained by the HMM-based local network alignment method using the global node correspondence scores and the individual node similarity scores for $L = 10$ (A,D), $L = 20$ (B,E), and $L = 30$ (C,F).

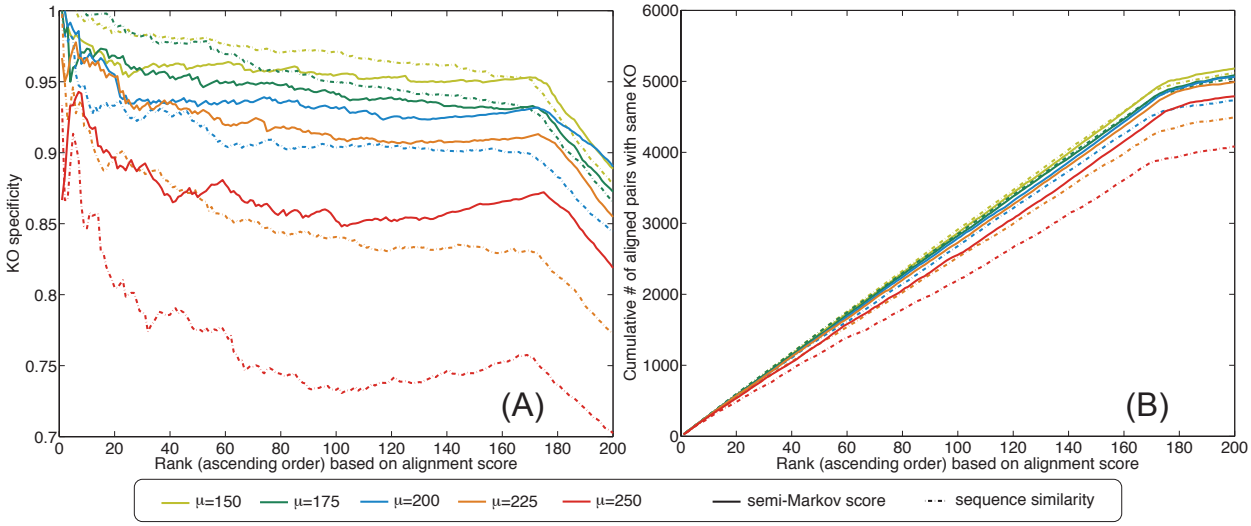

Figure 3: Functional specificity (A) and coverage (B) of the synthetic network alignment obtained using global correspondence scores and individual node similarity scores for various $\mu = \{150, 175, 200, 225, 250\}$.

more critical to utilize topological information to identify orthologs. This can be seen when $\mu = 250$, in which case the use of global correspondence scores can remarkably improve the accuracy of the alignment. It is also apparent from the figure that the performance of local network alignment using individual node similarity degrades much faster as the individual similarity gets less informative. This implies that we can obtain more robust and reliable alignment results by incorporating the global correspondence scores into the HMM-based local network alignment algorithm. (The URL for downloading the synthetic networks used in our experiments can be found in the *Supplementary materials*.)

Aligning microbial PPI networks

For further evaluation of the proposed method, we performed pairwise alignments of three microbial PPI networks obtained from [7]. In our experiments, we aligned the *E. coli* network and the *C. crescentus* networks to detect conserved functional modules in the two networks. Similarly, we also performed a pairwise alignment between the *E. coli* and the *S. typhimurium* networks to find conserved modules in these networks. As before, we have assessed the accuracy of the alignment results using two metrics—namely, specificity and coverage—based on the KEGG ortholog (KO) group annotations [25] of the proteins in the microbial networks. A protein alignment is regarded as being correct if the aligned

proteins have the same KO group annotations, and incorrect if the annotations do not agree.

In these experiments, the parameters of the HMMs were chosen as follows. First, the transition scores $t_w(u_i, u_j)$ in (1) were determined based on the presence (or absence) of protein interactions in the microbial networks determined by the SRINI algorithm [27]:

$$t_w(u_i, u_j) = \begin{cases} 0, & \text{if interaction between } u_i \text{ and } u_j \text{ exists;} \\ -\infty, & \text{otherwise.} \end{cases} \tag{5}$$

Second, we used the BLASTP hit scores between protein pairs, provided in [7], as the individual node similarity scores. The global correspondence scores were computed according to the semi-Markov random walk approach described earlier. These two types of node similarity scores were normalized such that they lie in the same range.

Based on the constructed HMMs, we used our local network alignment algorithm to find the 200 top-scoring pathway alignments with gaps. This experiment was also repeated for several different virtual path length: $L = 10, 20$, and 30. In all our experiments, we disallowed multiple occurrences of identical protein pairs in a given path alignment. The cumulative specificity cs_k of the pairwise alignment of the *E. coli* and the *C. crescentus* networks using the two different types of node similarities are shown in Figs. 4(A), (B), and (C) for $L = 10, 20$, and 30, respectively. The results of the pairwise alignment of the *E. coli* and the *S. typhimurium* networks are shown in Figs. 4(D), (E), and (F). Figure 5 shows the cumulative coverage cc_k of the predicted path alignments using the two different node similarity scores.

As we can see from the pairwise alignment results of the *E. coli* and the *C. crescentus* networks, shown in Fig. 4(A), (B), (C) and Fig. 5(A), (B), (C), when the coverage of the predicted path alignments is comparable, using the global correspondence scores results in higher specificity compared to using the individual node similarity scores. This implies that HMM-based network alignment based on global correspondence scores can more effectively capture the functional similarity between nodes. However, as we can see in Fig. 5(D), (E), (F), the protein pairs aligned using the semi-Markov random walk based global correspondence scores are less annotated (as reflected in the lower coverage cc_k) for the pairwise alignment of the *E. coli* and the *S. typhimurium* networks, in which case the specificity of the predicted alignment is not necessarily improved by the global scores. This can be seen in Fig. 4(D), (E), (F). One possible explanation for this observation is that the KO group annotations may have been curated largely baed on sequence similarity between proteins. For example, for remote orthologs that do not have high sequence

similarity, it may be practically difficult to judge to which KO group they should belong since there is not enough evidence. From this point of view, network alignment using global correspondence scores obtained from semi-Markov random walk could be used to validate and improve functional annotation of proteins.

In order to further examine the biological significance of the alignment results obtained from the proposed method, we retrieved the unannotated protein pairs that are aligned in the top path alignments predicted by comparing the *E. coli* and the *S. typhimurium* networks. As these proteins do not have curated functional annotations, such as GO terms or KO groups, we manually checked the protein information for the first 20 unannotated pairs in the Protein database at the National Center for Biotechnology Information (NCBI).[1] Table 1 shows the assigned gene names and regions names based on the GenInfo Identifiers (GIs) of these aligned proteins. Among these 20 protein pairs that are unannotated in the KEGG database, 18 pairs consist of proteins that have been assigned with the exactly same gene names and region names. In fact, these protein pairs indeed have similar cellular functionalities, where many of them are membrane proteins. Further information about these protein pairs can be found at the URLs included in the *Supplementary materials*. For the remaining two protein pairs (shown in bold face in Table 1), we could see that the aligned proteins were not assigned with the same gene names because the proteins in the *S. typhimurium* network were not annotated. When we checked the region names of the aligned proteins in each pair, we could see that the proteins in the first pair have the same region name "Transposase_31" and those in the second pair have the same region name "DUF1131". These observations suggest that the HMM-based network alignment method using semi-Markov random walk scores may provide a promising framework for automatic functional annotation of proteins.

Conclusion

In this paper, we studied the effect of using a global similarity scoring scheme to measure the node similarity and incorporating these global scores in the HMM-based local network alignment algorithm. We used the semi-Markov random walk framework to compute the global correspondence scores between nodes in different networks. The resulting scores can effectively combine the topological similarity of the subnetworks around the network nodes as well as their individual molecular similarity. Experimental results on microbial protein-protein interaction networks and synthetic scale-free networks show that the use of global correspondence scores can better identify paths with similar topological properties, thereby

[1] `http://www.ncbi.nlm.nih.gov/protein/`

Table 1: Gene names and Region names based on the GenInfo Identifiers (GIs) of the top 20 unannotated protein pairs that are aligned in the top conserved paths. Synonymous gene names are shown within parentheses.

E. coli			*S. typhimurium*		
GI	Gene name	Region name	GI	Gene name	Region name
16131641	wzzE	Wzz	16767194	wzzE	Wzz
49176398	viaA (yieM)	VWA_YIEM_type	39546380	yieM	VWA_YIEM_type
16131399	yhjJ	PqqL	16766899	yhjJ	PqqL
16131130	aaeB (yhcP)	FUSC	16766659	yhcP	FUSC
16130240	**yfcI**	**Transposase_31**	16767050	**STM3766**	**Transposase_31**
49176226	bamC (nlpB)	Lipoprotein_18	16765808	nlpB	Lipoprotein_18
16129342	ydbH	DctA-YdbH	16764990	ydbH	DctA-YdbH
49176233	sseB	SseB	16765855	sseB	SseB
16129572	ydgA	PRK11367	16764812	ydgA	PRK11367
16131557	yidR	propeller_TolB	16767096	yidR	TolB
16131404	bcsB (yhjN)	BcsB	16766904	yhjN	BcsB
16130950	ygiF	CYTH-like_Pase_CHAD	16766502	ygiF	CYTH-like_Pase_CHAD
16131855	yjbH	DUF940	16767475	yjbH	DUF940
16128005	yaaW (htgA)	Ubiq_cyt_C_chap	16763400	htgA	Ubiq_cyt_C_chap
16130391	ypfG	DUF1176	16765796	ypfG	DUF1176
16130357	**yfeY**	**DUF1131**	16765767	**STM2447**	**DUF1131**
49176330	yhdP	PRK10899	16766664	yhdP	PRK10899
16131526	yicH	AsmA	16767033	yicH	AsmA
16131275	yrfF	IgaA	16766783	yrfF	IgaA
16129282	ycjX	DUF463	16765028	ycjX	DUF463

improving the specificity of the predicted alignment. We believe that the proposed alignment scheme can provide an effective and computationally efficient framework for developing robust and accurate functional annotation tools for proteins.

Authors contributions

Conceived and designed the experiments: XQ, SMES, BJY. Performed the network alignment experiments: XQ. Implemented the semi-Markov random walk based scoring scheme: SMES. Analyzed the data and wrote the paper: XQ, SMES, BJY.

References

1. Osman A: **Yeast two-hybrid assay for studying protein-protein interactions**. *Methods Mol. Biol.* 2004, **270**:403–422.
2. Aebersold R, Mann M: **Mass spectrometry-based proteomics**. *Nature* 2003, **422**:198–207.
3. Kelley BP, Sharan R, Karp RM, Sittler T, Root DE, Stockwell BR, Ideker T: **Conserved pathways within bacteria and yeast as revealed by global protein network alignment**. *Proc. Natl. Acad. Sci. U.S.A.* 2003, **100**:11394–11399.
4. Koyut*ü*rk M, Grama A, Szpankowski W: **An efficient algorithm for detecting frequent subgraphs in biological networks**. *Bioinformatics* 2004, **20**:SI200–207.
5. Pinter R, Rokhlenko O, Yeger-Lotem E, Ziv-Ukelson M: **Alignment of metabolic pathways**. *Bioinformatics* 2005, **21**(16):3401–3408.
6. Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T: **Conserved patterns of protein interaction in multiple species**. *Proc. Natl. Acad. Sci. U.S.A.* 2005, **102**:1974–1979.
7. Flannick J, Novak A, Srinivasan B, McAdams H, Batzoglou S: **Græmlin: general and robust alignment of multiple large interaction networks**. *Genome Res* 2006, **16**(9):1169–1181.
8. Li Z, Zhang S, Wang Y, Zhang X, Chen L: **Alignment of molecular networks by integer quadratic programming**. *Bioinformatics* 2007, **23**(13):1631–1639.
9. Yang Q, Sze S: **Path matching and graph matching in biological networks**. *J Comput Biol* 2007, **14**:56–67.
10. Singh R, Xu J, Berger B: **Global alignment of multiple protein interaction networks with application to functional orthology detection**. *Proc. Natl. Acad. Sci. U.S.A.* 2008, **105**:12763–12768.
11. Flannick J, Novak A, Do CB, Srinivasan BS, Batzoglou S: **Automatic parameter learning for multiple local network alignment**. *J. Comput. Biol.* 2009, **16**:1001–1022.
12. Klau G: **A new graph-based method for pairwise global network alignment**. *BMC Bioinformatics* 2009, **10**(Suppl 1):S59.
13. Liao CS, Lu K, Baym M, Singh R, Berger B: **IsoRankN: Spectral methods for global alignment of multiple protein networks**. *Bioinformatics* 2009, **25**:i253–258.
14. Qian X, Sze SH, Yoon BJ: **Querying pathways in protein interaction networks based on hidden Markov models**. *Journal of Computational Biology* 2009, **16**:145–157.
15. Qian X, Yoon BJ: **Effective identification of conserved pathways in biological networks using hidden Markov models**. *PLoS ONE* 2009, **4**:e8070.
16. Tian W, Samatova N: **Pairwise alignment of interaction networks by fast identification of maximal conserved patterns**. In *Pac Symp Biocomput, Volume 14* 2009:99–110.

17. Zaslavskiy M, Bach F, Vert J: **Global alignment of protein-protein interaction networks by graph matching methods**. *Bioinformatics* 2009, **25**:259–267.

18. Sahraeian SME, Yoon BJ: **Fast network querying algorithm for searching large-scale biological networks**. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* 2011.

19. Sharan R, Ideker T: **Modeling cellular machinery through biological network comparison**. *Nat. Biotechnol.* 2006, **24**:427–433.

20. Qian X, Yoon BJ: **Comparative analysis of protein interaction networks reveals that conserved pathways are susceptible to HIV-1 interception**. *BMC Bioinformatics* 2011, **Suppl 1**(S19).

21. Sahraeian S, Yoon BJ: **A novel low-complexity HMM similarity measure**. *IEEE Signal Processing Letters* 2011, **18**(2):87–90.

22. Yoon BJ, Qian X, Sahraeian S: **Comparative analysis of biological networks using Markov chains and hidden Markov models**. *IEEE Signal Processing Magzines* 2011, **submitted**.

23. Barabasi AL, Albert R: **Emergence of scaling in random networks**. *Science* 1999, **286**:509–512.

24. Ashburner M, Ball C, Blake J, Botstein D, Butler H, et al.: **Gene Ontology: Tool for the unification of biology. the gene ontology consortium**. *Nat Genet* 2000, **25**:25–29.

25. Kanehisa M, Goto S: **KEGG: Kyoto encyclopedia of genes and genomes**. *Nucleic Acids Res* 2000, **28**:27–30.

26. Barabasi AL, Oltvai ZN: **Network biology: understanding the cell's functional organization**. *Nat. Rev. Genet.* 2004, **5**:101–113.

27. Srinivasan B, Novak A, Flannick J, Batzoglou S, McAdams H: **Integrated protein interaction networks for 11 microbes**. In *Proc of the 10th Annu Int Conf Res Comput Mol Bio (RECOMB 2006)* 2006.

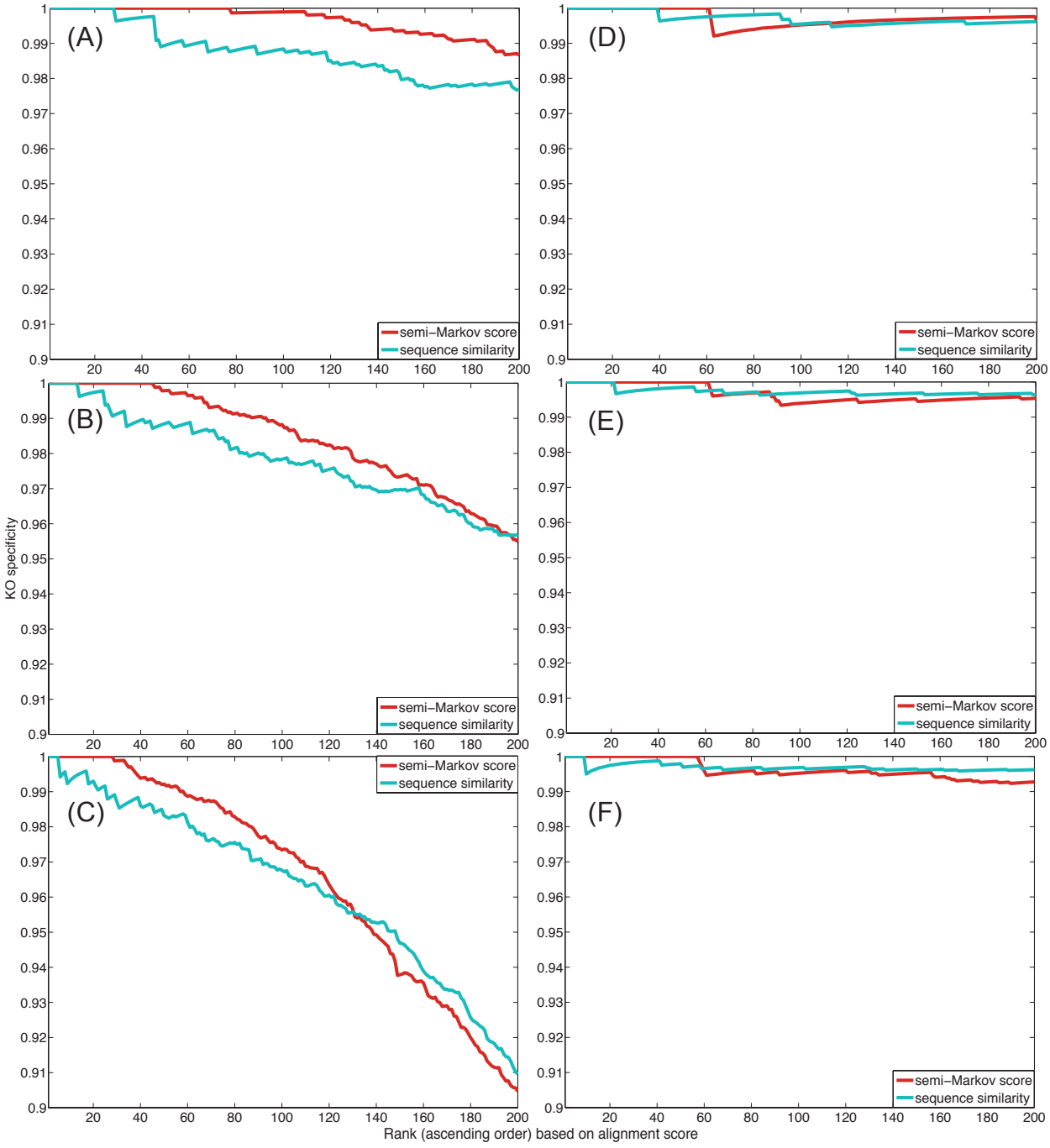

Figure 4: Functional specificity of the microbial network alignment obtained by the HMM-based local network alignment method using the global correspondence scores and the individual node similarity scores. The cumulative specificity of the top 200 path alignments are shown: (A) Pairwise alignment between *E. coli* and *C. crescentus* networks with $L = 10$; (B) Pairwise alignment between *E. coli* and *C. crescentus* networks with $L = 20$; (C) Pairwise alignment between *E. coli* and *C. crescentus* networks with $L = 30$; (D) Pairwise alignment between *E. coli* and *S. typhimurium* networks with $L = 10$; (E) Pairwise alignment between *E. coli* and *S. typhimurium* networks with $L = 20$; (F) Pairwise alignment between *E. coli* and *S. typhimurium* networks with $L = 30$.

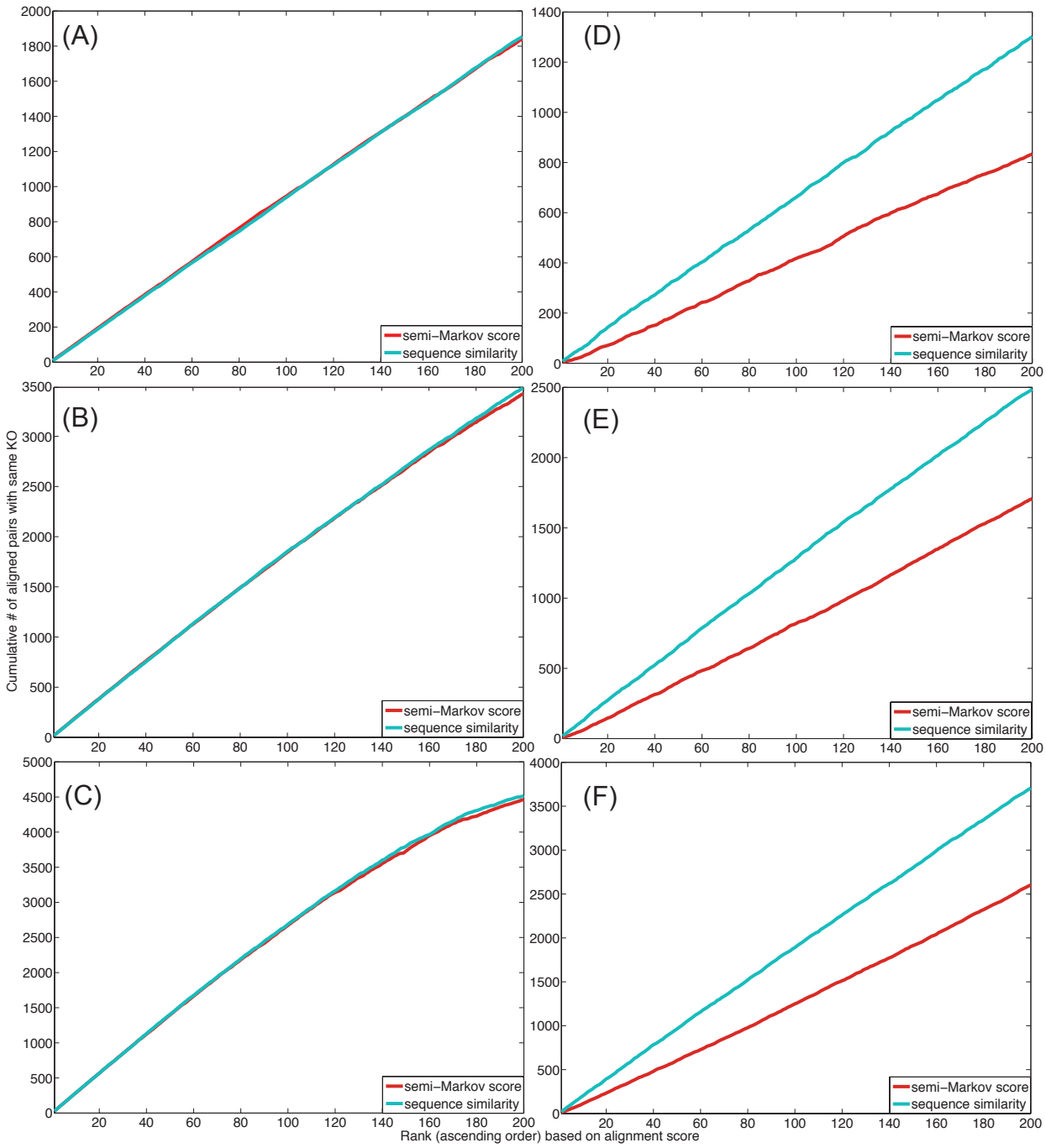

Figure 5: Functional coverage for microbial network alignment using the new semi-Markov node similarity and the original sequence similarity: The cumulative sensitivity of the top 200 aligned pathways obtained from (A) the pairwise alignment between *E. coli* and *C. crescentus* networks with $L = 10$; (B) the pairwise alignment between *E. coli* and *C. crescentus* networks with $L = 20$; (C) the pairwise alignment between *E. coli* and *C. crescentus* networks with $L = 30$; (D) the pairwise alignment between *E. coli* and *S. typhimurium* networks with $L = 10$; (D) the pairwise alignment between *E. coli* and *S. typhimurium* networks with $L = 20$; (D) the pairwise alignment between *E. coli* and *S. typhimurium* networks with $L = 30$.