Global Network Alignment Using Multiscale Spectral Signatures

Rob Patro and Carl Kingsford

December 14, 2011

Abstract

Motivation: Protein interaction networks provide an important system-level view of biological processes. One of the fundamental problems in biological network analysis is the global alignment of a pair of networks, which puts the proteins of one network into correspondence with the proteins of another network in a manner that conserves their interactions while respecting other evidence of their homology. By providing a mapping between the networks of different species, alignments can be used to inform hypotheses about the functions of unannotated proteins, the existence of unobserved interactions, the evolutionary divergence between the two species and the evolution of complexes and pathways.

Results: We introduce GHOST, a global pairwise network aligner that uses a novel spectral signature to measure topological similarity across disparate networks. It exhibits state-of-the-art performance on several network alignment tasks. We show that the spectral signature used by GHOST is highly discriminative, while the alignments it produces are also robust to experimental noise. When compared with other recent approaches, we find that GHOST is able to recover larger and biologically-significant, shared subnetworks between species.

Availability: An efficient and parallelized implementation of GHOST, released under the Apache 2.0 license, is available at
`http://cbcb.umd.edu/kingsford_group/ghost`
Contact: rob@cs.umd.edu

1. Introduction

We present a novel method for the global pairwise alignment of biological networks. Such alignments are crucial in analyzing the increasing amount of experimental data being generated by high-throughput techniques such as yeast two-hybrid screening [Fields and Song, 1989], tandem affinity purification mass spectrometry [Gavin et al., 2006], and chip-seq [Johnson et al., 2007] that reveal biological interactions within the cell.

A solution to the global network alignment problem is an injective mapping f from the nodes of one network G into another network H such that the structure of G is well preserved. This global mapping allows us to measure the similarity between proteins in G and those in H in terms of shared interaction patterns. By exposing large subnetworks with shared interactions patterns across species, a network alignment allows us to transfer protein function annotations from one organism to another using more information than can be captured by sequence alone. For example, it has been shown that, across species, the protein with the most similar sequence does not always play the same functional role [Sharan et al., 2005], and that topological information can be used to disambiguate sequence-similar proteins and determine functional orthology [Bandyopadhyay et al., 2006]. Additionally, by looking at the quality and magnitude of structure conserved between G and H, we can measure the similarity between these networks and infer phylogenetic relationships between the corresponding species [Kuchaiev and Pržulj, 2011]. We can also hypothesize the existence of unobserved interactions (missing edges), remove noise from error-prone, high-throughput experiments, and track the evolution of pathways.

Our approach to the global network alignment problem uses a novel measure of topological node similarity that is based on a multiscale spectral signatures. These signatures are composed from the spectra of the normalized Laplacian for subgraphs of varying sizes centered around a node. We combine this highly specific yet robust node signature with a seed-and-extend alignment strategy that explicitly enforces the proximity of aligned neighborhoods. We implement these ideas in our network alignment software, GHOST, which exceeds state-of-the-art accuracy under several different metrics of alignment quality.

There has been significant interest in the network alignment problem, and previous work can naturally be di-

vided into three main categories: approaches to local network alignment, approaches to network querying, and approaches to global network alignment. Because we are introducing a system for global network alignment, we restrict our discussion to the relevant work in this area.

Singh et al. [2008] introduced IsoRank that uses a recursively defined measure of topological similarity between nodes in different networks. They proposed an eigenvector-based formulation to discover a high-scoring matching. Liao et al. [2009] developed IsoRankN, which extends IsoRank with a new algorithm for multiple network alignment based on spectral clustering. The Graemlin aligner was originally developed by Flannick et al. [2006] to discover evolutionarily conserved modules across multiple biological networks. Later, it was extended [Flannick et al., 2009] to perform global multiple network alignment. However, this approach relies on a variety of additional information about the networks being aligned, including phylogenetic information. Further, sample alignments are required for the parameter learning phase of Graemlin2.

Recently, multiple attempts have been made to tackle the biological network alignment problem using graph matching. Klau [2009] introduced a non-linear integer program to maximize a structural matching score between two given networks, and then showed how the problem can be linearized, yielding an integer linear program (ILP), and finally suggested a Lagrangian relaxation approach to the ILP. The HopeMap approach of Tian and Samatova [2009] used an algorithm that iteratively merges conserved connected components. Zaslavskiy et al. [2009] explore the use of a number of graph matching methods, particularly the PATH and GA methods, which attempt to find a permutation matrix between vertices of the networks being aligned that maximizes a score that is a combination of the structural similarity and conserved interactions of the matched vertices. This optimization is NP-hard and they must rely on a relaxation to discover an approximate solution. Many similar graph-matching approaches have been applied to shape matching in computer graphics and computer vision [Torresani et al., 2008, Noma and Cesar, 2010, Duchenne et al., 2011]. All of these matching-based approaches require a large number of constraints to be placed on the set of potential alignments, usually in the form of homology information between the proteins of the networks being aligned. These constraints vastly reduce the search space and help bring these computationally burdensome methods into the realm of tractability. However, the hard constraints introduced by the homology information can have a negative effect on the ability of these methods to discover truly novel functional homologs between highly divergent species. In a way, these methods focus more on discovering conserved patterns of interactions between proteins that are already posited to be homologous, rather than on performing a truly *de novo* and unconstrained alignment of biological networks that is merely guided by homology information.

The GRAAL family of approaches, like IsoRank, performs truly unconstrained and *de novo* global pairwise alignments of biological networks. Kuchaiev et al. [2010] originally introduced GRAAL, which measures the topological similarity of nodes in different networks based on the distance between their graphlet degree signatures and aligns the networks using a seed-and-extend strategy. Milenkoviç et al. [2010] then introduced H-GRAAL, which relies on the same graphlet degree signatures used by GRAAL but performs the alignment of the networks by solving the linear assignment problem via the Hungarian algorithm [Kuhn, 1955] to discover the assignment that maximizes their score function. Finally, Kuchaiev and Pržulj [2011] introduced MI-GRAAL, which combines these two alignment strategies. It relies on a seed-and-extend alignment procedure but uses the Hungarian algorithm only to compute the assignment between local neighborhoods of the two graphs that maximizes the sum of their linear scoring function. MI-GRAAL also incorporates a number of other topological metrics, in addition to the graphlet degree signatures, to help quantify the topological similarity between nodes.

In this work, we will compare our results to those of multiple GRAAL aligners (MI-GRAAL is, in general, the best performing among this family of aligners) and IsoRank, which we view to be the most directly comparable with our work. This is because these methods place no *a priori* constraints on the networks being aligned. Although they can incorporate sequence similarity and other homology evidence into their scoring functions, they do not require the introduction of any hard constraints that limit the potential alignments and significantly reduce the space of considered solutions. It is possible that constrained graph-matching-based aligners and unconstrained *de novo* aligners are best suited to answer different questions and to be used in different scenarios, depending upon similarity of the considered organisms or the quantity and confidence of existing homology evidence. We will be concerned with the performance of GHOST as an unconstrained *de novo* aligner.

2. Methods

2.1 Spectral Signature

One of the primary contributions of our work is the introduction of a novel topological signature for nodes in a network. We use these signatures to guide our network alignment and to provide a measure of the similarity, or topological context, of nodes within their respective networks. Useful topological signatures should be precise, robust to topological variation, and fast to compute. Spectral graph theory provides tools that allow us to develop a signature having all of these properties.

There is a well-studied and strong relationship between the structure of a graph and the spectrum of its adjacency matrix and other related matrices. For example, isomorphic graphs are necessarily cospectral, though cospectral graphs are not necessarily isomorphic. However, simple comparison of spectra provide a powerful isomorphism filter in practice. In fact, using the eigenvalues and associated eigenvectors of graphs, Babai et al. [1982] developed an algorithm to decide graph isomorphism that is polynomial in the algebraic multiplicity of the graph.

The spectra of graphs are also robust to topological variations. Wilson and Zhu [2008] show that the distance between the spectra of the normalized Laplacian of graphs correlates well, at least for small perturbations, with the true edit distance between the graphs. Further, such spectra are efficient to compute. It takes $O(n^3)$ time to compute the spectrum for dense graphs with n vertices. However, for sparse graphs, like the biological graphs in which we are interested, faster algorithms exist [Pan and Chen, 1999]. For any subgraph, the computation of the spectrum is an independent operation and can be parallelized.

Our vertex signature is based on the spectrum of the normalized Laplacian for subgraphs of certain radii centered around a vertex. Consider a graph $G = (V, E)$ and vertex v. We denote by G_v^k the induced subgraph on all nodes whose unweighted shortest path length from v is less than or equal to k. We denote by W_v^k the weighted adjacency matrix of G_v^k. Finally, let the matrix D_v^k be given by

$$D_v^k[i,j] = \begin{cases} \sum_{k=1}^{n} W_v^k[i,k] & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Then, the normalized Laplacian of G_v^k is given by $\mathcal{L}_v^k = (D_v^k)^{\frac{1}{2}}(I - W_v^k)(D_v^k)^{\frac{1}{2}}$, where I is the appropriately-sized identity matrix. The eigendecomposition of this normalized Laplacian yields $\mathcal{L}_v^k V = \Lambda V$, where the sizes of V and Λ are the same as that of $\mathcal{L}_v^k$, but Λ is a diagonal matrix. We denote spectrum of $\mathcal{L}_v^k$ by $\sigma(\mathcal{L}_v^k)$, which is simply the entries along the main diagonal of Λ.

Many properties of $\sigma(\mathcal{L}_v^k)$ make it an enticing candidate for a vertex signature. Since the $\mathcal{L}_v^k$ is a positive, symmetric, semi-definite matrix with real entries, $\sigma(\mathcal{L}_v^k)$ consists entirely of non-negative real numbers. Further, the entries of $\sigma(\mathcal{L}_v^k)$ are bounded below by 0 and above by 2. Finally, many topological properties of a graph, such as the number of spanning trees, the Cheeger constant, and the distribution of path lengths are known to be related to the spectrum of its Laplacian [Chung, 1997].

However, for different vertices, the size of their k-hop neighborhoods will vary and thus the length of their spectra will be different and so the spectra cannot be directly compared. To overcome this difficulty, we consider the histogram of each spectrum instead, effectively computing a density estimate of the spectrum and discretizing the result. This yields a commensurate signature for each G_v^k, which we will denote as $\mathcal{S}_v^k$.

To compare the topological context of vertices at different scales, we simply consider the induced subgraphs for a range of different radii centered about v (i.e. $G_v^1, G_v^2, \ldots, G_v^k$). This leads, in turn, to a set of different spectra, and subsequently, different signatures. However, since the radii have the same meaning across different vertices and graphs (it is just the diameter of the neighborhood), the corresponding signatures can be compared directly and independently of the signatures at other radii. This leads to a simple scheme for comparing the topological contexts of two vertices at multiple scales using our signature. Consider two graphs, $G = (V_G, E_G)$ and $H = (V_H, E_H)$, with $u \in V_G$ and $v \in V_H$, and a sequence of radii $R = [1, 2, \ldots, k]$. We compute the distance between the signatures of u and v for this sequence of radii as

$$D_{\text{topo}}(\mathcal{S}_u^R, \mathcal{S}_v^R) = \sum_{r \in R} d(\mathcal{S}_u^r, \mathcal{S}_v^r), \quad (2)$$

where $d(\cdot, \cdot)$ can be any desired histogram distance. Currently, we simply consider the ℓ_1 norm — that is $d(x, y) = ||x - y||_1$ — but other distance measures such as Earth Movers Distance [Rubner et al., 1998], or the Quadratic Chi [Pele and Werman, 2010] distance may work well.

In a manner similar to IsoRank [Singh et al., 2008], we can incorporate sequence information into our distance measure between two proteins u and v by using a simple combination of the topological distance — $D_{\text{topo}}(\mathcal{S}_u^R, \mathcal{S}_v^R)$ as defined in Equation (2) — and a sequence distance, $D_{\text{seq}}(u, v)$, such as the symmetrized

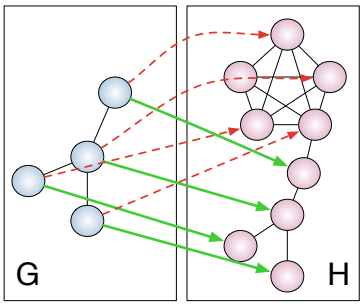


Figure 1: *The mapping from G to H given by the solid green arrows can be considered a better alignment than that given by the dashed red arrows, despite the fact that they both have the same edge correctness.*

BLAST E-value. The total distance measure is a linear combination of the topological and sequence distance, parameterized by some weight α and is given by

$$D_\alpha(u,v) = \alpha D_{\text{topo}}(\mathcal{S}_u^R, \mathcal{S}_v^R) + (1.0-\alpha) D_{\text{seq}}(u,v). \quad (3)$$

2.2 Measuring Alignment Quality

It is challenging to state the global network alignment problem formally and precisely because a "good" alignment balances two, often disparate, goals. A high-quality global alignment between two biological networks should reveal shared topological structure between the networks being aligned, while also respecting the strong evidence for homology revealed via sequence analysis.

Neither of these goals, however, should act as hard constraints when aligning two networks and a high-quality global network alignment should strive to satisfy both the topological and sequence requirements. This naturally leads to two distinct measures for the quality of network alignments; one measures *topological quality*, the degree of shared structure revealed between the two networks, and the other measures *biological quality*, how well the alignment respects the biological and functional similarities of the proteins.

Topological Quality. A topological quality metric should measure the degree to which the structure of G is preserved, under f, when mapped into H. For example, we expect that an alignment of high topological quality will map interacting proteins in G to interacting proteins in H. The most common measure of topological quality is edge correctness, which measures the percentage of edges from G which are aligned to edges in H. Let $G[V]$ be the induced subgraph of G on the vertex set V, $f(V) = \{f(v) \mid v \in V\}$, $f(E) = \{(f(u), f(v)) \mid (u,v) \in E\}$ and $f(G) = (f(V_G), f(E_G))$. Then, given the networks G and H and the alignment f, the edge correctness (EC) is defined as

$$\text{EC}(G,H,f) = \frac{\left|f(E_G) \cap E_H\right|}{|E_G|}. \quad (4)$$

Despite its prevalence, edge correctness fails to differentiate alignments that one might intuitively consider to be of different topological quality (see Figure 1) because it accounts only for the number of edges from G that are mapped into H and incorporates no notion of the similarity between G and the induced subgraph of $f(G)$.

We introduce a new measure of topological quality, the induced conserved structure (ICS) score, that incorporates a richer notion of conserved structure than EC. We define the ICS score between G and H induced by the alignment f as

$$\text{ICS}(G,H,f) = \frac{\left|f(E_G) \cap E_H\right|}{|E_{H[f(V_G)]}|}. \quad (5)$$

Notice that, for the example given in Figure 1, while the edge EC score of both the green and red mappings is 1, the ICS successfully distinguishes the two cases. In particular, the ICS of the green mapping remains 1, while the ICS of the red mapping becomes 0.4, agreeing with the intuition that the green mapping conserves more structure than does the red mapping. Also, note that the ICS score is 1 if and only if G is isomorphic to $H[f(V_G)]$. Thus, alignments that map subgraphs of G into denser subgraphs of H, where there are potentially many more mappings, will be punished under the ICS score while they will not be punished under the standard edge correctness score.

Another common measure of the topological quality of an alignment that we consider is the size of the largest connected shared component (LCSC). A single, large, connected shared component is better than a collection of many small shared components because it represents a larger and more coherent shared structure. We expect that an alignment of higher topological quality will find a larger shared structure between the two networks than will an alignment of lower quality.

Biological Quality. Given an alignment, $f : G \to H$, a measure of biological quality should evaluate the similarity of p and $f(p)$ in terms of biological function. The most common measure of similarity computes the enrichment

of shared Gene Ontology [The Gene Ontology Consortium et al., 2000] (GO) annotations between the mapped proteins. The greater the enrichment, the higher the biological quality of the alignment. In previous work [Singh et al., 2008, Kuchaiev and Pržulj, 2011], two GO annotations are considered the same only if they are identical.

This metric has two main disadvantages. First, many GO terms are assigned largely based on sequence homology to proteins with verified annotations, which strongly biases the results in favor of alignments that ignore topology completely and align proteins based solely on sequence similarity. Additionally, measuring the functional enrichment between proteins by considering only exact overlap between their associated GO annotations ignores the hierarchical structure of annotation similarity encoded in the ontology. Though previous work [Singh et al., 2008, Liao et al., 2009, Kuchaiev et al., 2010, Kuchaiev and Pržulj, 2011] considers only this exact overlap metric, it is almost certainly misleading.

While the first issue remains a concern, we address the latter by using an additional metric of protein function similarity that takes into account the relationships between annotations encoded by the GO hierarchy. Pesquita et al. [2009] recently compared a number of methods for computing protein similarities based on GO annotations. They find that one of the best performing methods computes the similarity of GO terms using the Resnik ontological similarity measure and combines annotation similarities using the best-match average strategy to obtain a functional similarity measure on proteins. We adopted an implementation of this measure provided in the csbl.go R-project package [Ovaska et al., 2008]. We denote this similarity measure by $s_a(p_1, p_2)$, where a is an aspect — Biological Process (BP), Molecular Function (MF) and Cellular Component (CC) — of GO. The similarity measure between networks G and H induced by the alignment f under the GO aspect a is given by $s_a(G, H, f) = \frac{1}{|V_G|} \sum_{p \in V_G} s_a(p, f(p))$.

3. Alignment Procedure

Given a distance function such as D_α described by Equation (3), there are many ways one can go about computing an alignment. The most straightforward approaches to computing an alignment are solving the linear assignment problem (LAP), approximating a solution to the quadratic assignment problem (QAP), or using a seed-and-extend strategy. For example, IsoRank approximates a solution to a modified QAP problem using an eigenvector-based approach while MI-GRAAL combines a seed-and-extend and LAP strategy. GHOST supports all 3 of these approaches. We show results for GHOST's default mode, which uses a seed-and-extend strategy combined with a QAP procedure to align local neighborhoods as described below.

Much like the strategy used in the sequence alignment tool `BLAST` [Altschul et al., 1990], the seed-and-extend strategy employed by GHOST seeds regions of an alignment with high scoring pairs of nodes from the different networks and then extends the alignments around the neighborhoods of these two nodes. This procedure executes in rounds until all nodes from the smaller of the two networks have been aligned with some node from the larger network.

First, an alignment is seeded with a high-scoring match $\hat{M}^0 = (\hat{M}^0_G, \hat{M}^0_H)$. This is a pair of vertices between which the specified D_α is minimal. Then, we consider all pairwise matches between the 1-hop neighborhoods of these two vertices, $M = \left[(i, j) \mid i \in \mathcal{N}(\hat{M}^0_G)), j \in \mathcal{N}(\hat{M}^0_H))\right]$, and form a quadratic assignment matrix Q given by:

$$Q[a, b] = \begin{cases} 1 - D_\alpha(M[a][0], M[a][1]) & \text{if } a = b \\ C(M[a], M[b]) & \text{otherwise} \end{cases}$$

$M[a]$ is the a^{th} match in M, and $C(M[a], M[b]) = \exp(\frac{-|d(a,b,0)-d(a,b,1)|}{d(a,b,0)+d(a,b,1)})$ measures the pairwise consistency of potential matches $M[a]$ and $M[b]$, where $d(a, b, 0) = D_{\text{topo}}(M[a][0], M[b][0])$ and $d(a, b, 1) = D_{\text{topo}}(M[a][1], M[b][1])$. Solving the spectral relaxation of the quadratic assignment problem [Leordeanu and Hebert, 2005] gives an approximate solution, s, which assigns a confidence, $\text{Conf}_s(\cdot, \cdot)$, to each potential match (that is, to each $m \in M$). The matches are then ranked by confidence and the top pair is aligned and becomes the seed for the next round. We continue extending the alignment in this manner, covering larger topological neighborhoods of the original seed nodes, until no more nodes can be aligned. Then, the next seed pair, $\hat{M}^1$, is chosen from among the unaligned nodes and the same procedure is applied to extend the alignment around this seed. This process continues until all nodes from V_G (assumed, w.l.o.g., to be smaller than V_H) have been aligned. This process is given more formally in Algorithms 1 and 2.

Algorithm 1: SeedAndExtend

input : Networks G and H
output: Alignment f

$P \leftarrow \{\}$; *// Initialize (min) heap*
$f \leftarrow \{\}$; *// Initialize empty alignment*
foreach $(x, y) \in V_G \times V_P$ **do**
  push$(P,\ (x, y, D_\alpha(x, y)))$;
while P *is not empty* **do**
  $(t_G, t_H) \leftarrow$ **pop**(P);
  if t_G *and* t_H *are not already aligned* **then**
    GreedyQAPExtend$(G, P, (t_G, t_H), f)$;
return f

Algorithm 2: GreedyQAPExtend

input : Networks G and H, seed pair (u_G, u_H), current alignment f
side-effect: f extended with some neighbors of u_G, u_H

$P \leftarrow \{(u_G, u_H)\}$; *// Initialize (max) heap*
while P *is not empty* **do**
  $(t_G, t_H) \leftarrow$ **pop**(P);
  if t_G *and* t_H *are not already aligned* **then**
    // Align neighborhoods using the approximate
    // quadratic assignment procedure, QA
    $s \leftarrow QA(\mathcal{N}(t_G), \mathcal{N}(t_H))$;
    foreach $(x, y) \in s \setminus (f(G) \times f(H))$ **do**
      push$(P,\ (x, y, \mathrm{Conf}_s(x, y)))$;
      $f(x) \leftarrow y$;

4. Results

We evaluated the performance of GHOST in a number of different scenarios. First, we consider two tests that have been used in the past to assess topological alignment quality. These tests, self-alignment and self-alignment with noise, are instructive because the correct node mapping is known when aligning a network to itself. This allows us to measure accuracy in a way that is not possible when comparing networks from different species. The results of these experiments provide important evidence about the robustness and specificity of different topological signatures and the ability of different global alignment approaches to align two networks based solely on topological information. Subsequently, we consider the alignment between high-confidence protein-protein interaction networks of a pair of bacteria and a pair of eukaryotes. Here, we use the metrics in Section 5.2 to measure the topological and biological quality of our alignments.

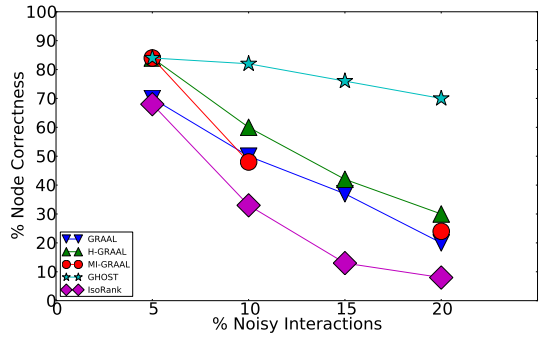
4.1 Self-Alignment

For networks with many similar sub-regions, even a self-alignment in the absence of noise can be difficult. To demonstrate this difficulty, we consider a self-alignment of the largest connected component of a high-confidence network of the bacterium *Mesorhizobium loti* (*M. loti*). This network was obtained from the interactions reported in the study by Shimoda et al. [2008] and consists of 3006 interactions among 1655 proteins. The alignment produced by GHOST is an automorphism of the graph, with an edge correctness of 100% and a node correctness (the fraction of nodes which were aligned with themselves) of 79%. The alignment produced by IsoRank had an edge correctness of 76% and a node correctness of 53%, while the alignment produced by MI-GRAAL had a edge correctness of 38% and node correctness of only 0.3%. Because MI-GRAAL is probabilistic in nature, we performed this alignment multiple times, using a wide variety and combination of the topological features suggested in Kuchaiev and Pržulj [2011], to ensure that this failure of self-alignment was not coincidental. None of these subsequent MI-GRAAL alignments differed in topological quality — either node or edge correctness — by more than a fraction of a percent. IsoRank produced an alignment of significantly higher topological quality than the one discovered by MI-GRAAL; this is different from what we see in the rest of the tests described below.

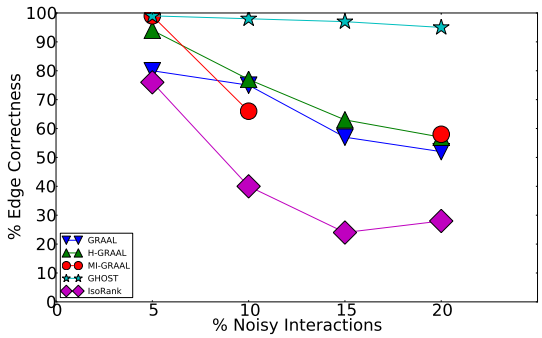
Despite the fact that its node correctness is only 79%, GHOST's alignment is structurally perfect. Without more information beyond what is provided by the network itself, one cannot hope to obtain a better alignment than the one produced by GHOST.

4.2 Self-Alignment Under Noise

We also re-performed the experiment originally carried out by Milenkoviç et al. [2010], where progressively noisier variants of the *S. cerevisiae* interaction network are aligned to the high-confidence network of Collins et al. [2007]. The higher noise networks are created by starting with the highest confidence network, and then adding interactions (constrained to the original, high confidence protein set) in decreasing order of experimental confidence. Since this is again a self-alignment, and sequence



(a) *Node Correctness*



(b) *Edge Correctness*

Figure 2: *Performance of various aligners on a noisy yeast PPI under the node and edge correctness metrics. Note: In the* 15% *noise case, the performance numbers of MI-GRAAL are not given because it failed to run to completion.*

information would allow the almost perfect identification of the correspondences between nodes, we consider a purely topological alignment (i.e. $\alpha = 1.0$). We explore how the fraction of correctly aligned nodes changes as larger quantities of noisy interactions are added to the high-confidence network (Figure 2).

In the case with the fewest noisy interactions, most of the programs achieve similar performance. However, as the number of noisy interactions increases, GHOST outperforms all of the other approaches by an increasing margin. By the time 20% of the noisy interactions have been included in the network, the node correctness of GHOST is more than twice that of the next-best-performing aligner, while the edge correctness is over 30% greater. There also seems to be a substantial gap between IsoRank and the rest of the alignment procedures in terms of both the node and edge correctness. This is indicative of a trend we observe when aligning real biological networks as well (see below), where the topological quality of the alignments produced by IsoRank, even with a large weight being placed on the topological score, seems to fall behind those produced by the other aligners.

The performance of GHOST in this set of experiments suggests that the spectral signature is robust to the presence of noise in the network, significantly more so than the graphlet degree signatures used in the GRAAL aligners. These results agree with existing evidence, such as that presented by Wilson and Zhu [2008], that the spectral distance between graphs is robust to small topological changes. Both this robustness, and the specificity of the spectra, seem to carry over to our topological signatures, and do not appear to be negatively affected by the discretization we perform to deal with graphs of different order.

4.3 Alignments Between Different Species

We performed an alignment of the high-confidence protein interaction networks of *Campylobacter jejuni* (*C. jejuni*) and *Escherichia Coli* (*E. coli*). Both of these bacterial species are well-studied model organisms. In order to draw the most appropriate comparisons to MI-GRAAL, we use the same versions of the interaction networks that were used by Kuchaiev and Pržulj [2011]. Thus, we considered *E. coli* network composed of interactions from the data of Peregrín-Alvarez et al. [2009], consisting of 1941 proteins among which there are 3989 interactions. We consider the *C. jejuni* network which consists of the high-confidence interaction from the data of Parrish et al. [2007], containing of 2988 interactions among 1111 proteins.

We consider multiple measures (as introduced in Section 5.2) of the topological and biological quality of the alignments produced by the different approaches. To calculate annotation enrichment, we rely on the set of GO annotations for each protein retrieved from the European Bioinformatics Institute website in June of 2011. To compute the GO similarities, we use the gene ontology retrieved on Nov. 10, 2011. When producing alignments using MI-GRAAL, we included graphlet degree signatures, clustering coefficients and sequence similarity scores — the topological features that Kuchaiev and Pržulj [2011] found to lead to the highest scoring and most stable alignments. When aligning two networks, MI-GRAAL determines the value of α — the parameter that trades off between functional and sequence similarity — which opti-

mizes its own scoring function, and so no α value was provided. For IsoRank and GHOST, we varied α between 0 and 1 in increments of 0.05. In our experiments, the biological quality of the IsoRank solutions varied little, so we report the alignment with the highest topological quality. For GHOST, we report scores for $\alpha = 0.25$, which seemed to produce alignments of high topological quality without significantly sacrificing biological quality.

Both GHOST and MI-GRAAL produce results of similar topological quality when aligning *E. coli* and *C. jejuni*; both achieve about 24% edge correctness. However, GHOST is able to uncover a slightly larger and denser connected shared component compared to MI-GRAAL, and also exhibits a slightly higher ICS score. Therefore, GHOST produces an alignment with a similar, and arguably better, topological quality than that of MI-GRAAL. The biological quality of GHOST's alignment is uniformly superior to MI-GRAAL's, consisting of a greater number of exactly overlapping GO annotations between aligned proteins, as well as a smaller overall s_{a} for all aspects of the gene ontology (see Tables 2 and 3). In particular, as we consider aligned protein pairs sharing multiple GO annotations, the difference between alignment of GHOST and MI-GRAAL becomes more pronounced.

The nature of IsoRank's alignment between these two networks is quite different than that of GHOST and MI-GRAAL. IsoRank (using $\alpha = 0.9$) produces an alignment that exhibits excellent biological quality — the highest GO term enrichment and the smallest d_{a} under the three aspects of the gene ontology. However, the topological quality of its alignment is substantially lower than that of the other aligners (Tables 1 to 3). Particularly striking is the size of the largest connected shared component discovered by IsoRank, which consists of only 12 nodes and 11 edges. This essentially tree-like component is $\sim$52 times smaller than the one discovered by GHOST and $\sim$48 times smaller than the one discovered by MI-GRAAL. This was the alignment of the highest topological quality discovered by IsoRank.

We also explored the ability of GHOST to align the protein interaction networks of distant eukaryotes by performing an alignment of the protein interaction networks of *Arabidopsis thaliana* and *Drosophila melanogaster*. We obtained the interactions for these networks from the HitPredict website [Patil et al., 2011]. HitPredict places interaction data for each species into three categories: high-confidence small-scale interactions (HCSS), high-confidence high-throughput interactions (HCHT), and low-confidence high-throughput interactions (LCHT). The high-confidence small-scale interactions are identified directly in small-scale experiments considering fewer than 100 interactions each. The HCHT interactions are those interactions identified in high-throughput experiments with a likelihood ratio greater than 1, or predicted from protein complex data. The low-confidence high-throughput interactions are those having a likelihood ratio less than 1. In our experiments, we considered only the high-confidence interactions — the union of those interactions in the HCSS and HCHT sets. This resulted in a network for *A. thaliana* having 2082 proteins and 4145 interactions. The *D. melanogaster* network consisted of 7615 interactions among 3792 different proteins.

The alignment quality is very similar to that of the bacterial networks (see Tables 1 to 3). The edge correctness and induced conserved structure scores of GHOST's and MI-GRAAL's alignments are both very similar. GHOST is again able to uncover a larger connected shared component — consisting of 1083 interactions among 1023 proteins — than the one found by MI-GRAAL — which has 1031 interactions among 976 proteins. However, the biological quality of GHOST's alignment are uniformly higher than that of MI-GRAAL's alignment. We again observe that the greater the number of shared GO annotations we consider, the larger the gap between the score of GHOST's and MI-GRAAL's alignment. Further, while the GO similarities under the biological process and molecular function aspects are statistically significant for the mappings produced by all three aligners, only IsoRank and GHOST exhibit a statistically significant alignment under the cellular component aspect for these two species.

Between these species, IsoRank produces an alignment with very low topological quality by all metrics. The largest common shared component consists of only 26 proteins and 26 interactions, which is $\sim$40 times smaller than the component discovered by GHOST. While IsoRank's alignment again gives the highest biological quality scores, the margin is significantly smaller this time, both with respect to the enrichment of shared GO terms between aligned proteins and the overall similarities between aligned proteins under all ontological aspects of the gene ontology. Perhaps because *A. thaliana* and *D. melanogaster* are more evolutionarily distant than are *C. jejuni* and *E. coli*, the homology evidence provided by sequence similarity is weaker. Since IsoRank seems to rely so heavily on sequence similarity, this could explain why the gap in the biological quality of the alignment is smaller between these organisms than between *C. jejuni* and *E. coli*.

4.4 Diversity of Alignments

We also compared the actual node mappings produced by GHOST, MI-GRAAL and IsoRank. In the *C. jejuni* vs. *E. coli* alignments, GHOST and MI-GRAAL share only 6.2% of aligned pairs, GHOST and IsoRank share 2.7% of aligned pairs and MI-GRAAL and IsoRank share 0.5% of aligned pairs. In the *A. thaliana* vs. *D. mel* alignments, GHOST and MI-GRAAL share 6.3% of aligned pairs, GHOST and IsoRank share 3.0% of aligned pairs and MI-GRAAL and IsoRank share 2.7% of aligned pairs. The low agreement between mappings is very surprising and suggest that many diverse alignments of similar topological and biological quality may exist and that more research needs to be carried out into informative metrics for distinguishing and measuring the quality of network alignments. Further, these results provide justification for exploring the local alignment problem, where an aligner would be allowed to expose multiple high-quality overlapping alignments.

5. Discussion

We have introduced GHOST, a novel framework for the global alignment of biological networks. At the heart of GHOST is a new spectral, multiscale node signature that we combine with a seed-and-extend approach to perform global network alignment. The spectral signature is highly discriminative and robust to small topological variations. We verify this robustness in Section 7.2, showing that GHOST outstrips the competition in aligning the *S. cerevisiae* protein interaction network to noisier variants of itself. In these experiments, as well as the self-alignment of the *M. loti* network, the accuracy of GHOST is significantly higher than that of either IsoRank or MI-GRAAL. These experiments are of particular interest, because the ground truth is known and the ability of different aligners to uncover shared topological structure can be accurately measured.

We find that the alignments produced by GHOST and MI-GRAAL when aligning the interaction networks of different species are of similar topological quality, although GHOST is able to discover a larger connected shared component. However, at a similar level of topological quality, the alignments produced by GHOST are of significantly better biological quality, specifically when we consider the enrichment of aligned proteins for multiple GO terms. Under all the metrics we consider, the alignments produced by GHOST and MI-GRAAL are of much higher topological quality than those produced by IsoRank. Even when very high weights are given to the structural score, IsoRank is unable to discover a significant portion of the shared topologies of the networks. Because IsoRank seems to strongly weight sequence similarity regardless of the α parameter we provide, it produces alignments which exhibit high biological quality according to the metrics we consider. Recall, however, the caveats inherent in the somewhat circular definition of biological quality (see Section 5.2). For example, for the *C. jejuni* and *E. coli* networks, a sequence-based alignment yields the highest GO term enrichment, significantly higher than even IsoRank's solution, but it uncovers very little shared topology.

Acknowledgments

The authors thank Jeremy Bellay, Geet Duggal, Darya Filippova, Justin Malin, Guillaume Marçais, Emre Sefer and Hao Wang for useful discussions.

Funding: This work was supported by the National Science Foundation [CCF-1053918, EF-0849899, and IIS-0812111]; the National Institutes of Health [1R21AI085376]; and a University of Maryland Institute for Advanced Studies New Frontiers Award.

References

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410, October 1990.

László Babai, D. Yu. Grigoryev, and David M. Mount. Isomorphism of graphs with bounded eigenvalue multiplicity. In *Proc. of the 14th Annual ACM Symposium on Theory of Computing*, STOC '82, pages 310–324, New York, NY, USA, 1982. ACM. ISBN 0-89791-070-2. doi: http://doi.acm.org/10.1145/800070.802206.

Sourav Bandyopadhyay, Roded Sharan, and Trey Ideker. Systematic identification of functional orthologs based on protein network comparison. *Genome Res.*, 16(3): 428–435, March 2006. ISSN 1088-9051. doi: 10.1101/gr.4526006.

F R K Chung. *Spectral Graph Theory*, volume 92. American Mathematical Society, 1997.

Sean R Collins, Patrick Kemmeren, Xue-Chu Zhao, Jack F Greenblatt, Forrest Spencer, Frank C P Holstege, Jonathan S Weissman, and Nevan J Krogan. Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Molecular Cellular Proteomics*, 6(3):439–450, 2007.

Olivier Duchenne, Francis Bach, In-So Kweon, and Jean Ponce. A tensor-based algorithm for high-order graph matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33 (12):2383–2395, 2011.

S. Fields and O. Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246, July 1989.

Jason Flannick, Antal Novak, Balaji S Srinivasan, Harley H McAdams, and Serafim Batzoglou. Graemlin: General and robust alignment of multiple large interaction networks. *Genome Res.*, 16(9):1169–1181, 2006.

Jason Flannick, Antal Novak, Chuong B Do, Balaji S Srinivasan, and Serafim Batzoglou. Automatic parameter learning for multiple local network alignment. *J. Computat. Biol.*, 16(8):1001–1022, 2009.

Anne-Claude Gavin, Patrick Aloy, Paola Grandi, Roland Krause, Markus Boesche, Martina Marzioch, Christina Rau, Lars Juhl Jensen, Sonja Bastuck, Birgit Dumpelfeld, Angela Edelmann, Marie-Anne Heurtier, Verena Hoffman, Christian Hoefert, Karin Klein, Manuela Hudak, Anne-Marie Michon, Malgorzata Schelder, Markus Schirle, Marita Remor, Tatjana Rudi, Sean Hooper, Andreas Bauer, Tewis Bouwmeester, Georg Casari, Gerard Drewes, Gitte Neubauer, Jens M. Rick, Bernhard Kuster, Peer Bork, Robert B. Russell, and Giulio Superti-Furga. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440 (7084):631–636, 03 2006.

David S. Johnson, Ali Mortazavi, Richard M. Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, 2007. doi: 10.1126/science.1141319.

Gunnar W Klau. A new graph-based method for pairwise global network alignment. *BMC Bioinformatics*, 10(Suppl 1):S59, 2009.

Oleksii Kuchaiev and Natasa Pržulj. Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics*, 27(6):1–7, 2011.

Oleksii Kuchaiev, Tijana Milenkovic, Vesna Memisevic, Wayne Hayes, and Natasa Pržulj. Topological network alignment uncovers biological function and phylogeny. *J. Royal Soc., Interface*, 7(50):1341–54, September 2010. ISSN 1742-5662. doi: 10.1098/rsif.2010.0063.

H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistic Quarterly*, 2:83–97, 1955.

M Leordeanu and M Hebert. A spectral technique for correspondence problems using pairwise constraints. *Tenth IEEE International Conference on Computer Vision ICCV05 Volume 1*, 2:1482–1489, 2005.

Chung-Shou Liao, Kanghao Lu, Michael Baym, Rohit Singh, and Bonnie Berger. IsoRankN: spectral methods for global alignment of multiple protein networks. *Bioinformatics*, 25(12):i253–i258, June 2009. ISSN 1460-2059.

Tijana Milenkoviç, Weng Leong Ng, Wayne Hayes, and Nataša Pržulj. Optimal network alignment with graphlet degree vectors. *Cancer Informatics*, 9:121–137, 2010.

A. Noma and R.M. Cesar. Sparse representations for efficient shape matching. In *Graphics, Patterns and Images (SIBGRAPI), 2010 23rd SIBGRAPI Conference on*, pages 186 –192, Sept. 2010.

Kristian Ovaska, Marko Laakso, and Sampsa Hautaniemi. Fast gene ontology based clustering for microarray experiments. *BioData mining*, 1(1):11, 2008.

Victor Y. Pan and Zhao Q. Chen. The complexity of the matrix eigenproblem. In *Proc. of the Thirty-first Annual ACM Symposium on Theory of Computing*, STOC '99, pages 507–516, New York, NY, USA, 1999. ACM. ISBN 1-58113-067-8.

Jodi R Parrish, Jingkai Yu, Guozhen Liu, Julie A Hines, Jason E Chan, Bernie A Mangiola, Huamei Zhang, Svetlana Pacifico, Farshad Fotouhi, Victor J DiRita, Trey Ideker, Phillip Andrews, and Russell L Finley. A proteome-wide protein interaction map for *Campylobacter jejuni*. *Genome Biol.*, 8(7):R130, 2007.

Ashwini Patil, Kenta Nakai, and Haruki Nakamura. HitPredict: a database of quality assessed proteinprotein interactions in nine species. *Nuc. Acids Res.*, 39 (Database issue):D744–D749, 2011.

Ofir Pele and Michael Werman. The quadratic-chi histogram distance family. In *ECCV*, 2010.

José M Peregrín-Alvarez, Xuejian Xiong, Chong Su, and John Parkinson. The modular organization of protein

interactions in *Escherichia coli. PLoS Computat. Biol.*, 5(10):e1000523, 2009.

Catia Pesquita, Daniel Faria, André O. Falcåo, Phillip Lord, and Francisco M. Couto. Semantic similarity in biomedical ontologies. *PLoS Computat. Biol.*, 5(7): e1000443, 07 2009.

Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *Computer Vision, 1998. Sixth International Conference on*, pages 59–66, 1998. doi: 10.1109/ICCV.1998.710701.

Roded Sharan, Silpa Suthram, Ryan M Kelley, Tanja Kuhn, Scott McCuine, Peter Uetz, Taylor Sittler, Richard M Karp, and Trey Ideker. Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. USA*, 102(6):1974–9, 2005. ISSN 0027-8424. doi: 10.1073/pnas.0409522102.

Yoshikazu Shimoda, Sayaka Shinpo, Mitsuyo Kohara, Yasukazu Nakamura, Satoshi Tabata, and Shusei Sato. A large scale analysis of proteinprotein interactions in the nitrogen-fixing bacterium *Mesorhizobium loti*. *DNA Research*, 15(1):13–23, 2008.

Rohit Singh, Jinbo Xu, and Bonnie Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proc. Natl. Acad. Sci. USA*, 105(35):12763–8, September 2008. ISSN 1091-6490. doi: 10.1073/pnas.0806627105.

The Gene Ontology Consortium, M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000. ISSN 1061-4036. doi: 10.1038/75556.

Wenhong Tian and Nagiza F Samatova. Pairwise alignment of interaction networks by fast identification of maximal conserved patterns. *Pacific Symposium On Biocomputing*, pages 99–110, 2009.

Lorenzo Torresani, Vladimir Kolmogorov, and Carsten Rother. Feature correspondence via graph matching: Models and global optimization. In *ECCV*, pages 596–609, 2008.

Richard C. Wilson and Ping Zhu. A study of graph spectra for comparing graphs and trees. *Pattern Recognition*, 41:2833–2841, 2008. doi: 10.1016/j.patcog.2008.03.011.

Mikhail Zaslavskiy, Francis Bach, and Jean-Philippe Vert. Global alignment of protein-protein interaction networks by graph matching methods. *Bioinformatics*, 25 (12):i259–1267, 2009.

Table 1: *The various topological quality metrics for various aligners on different networks. Columns* 3*,* 4 *and* 5 *list the edge correctness, induced conserved structure score, and the number of nodes and edges in the largest connected shared component discovered.*

Networks	Method	Edge Corr.	ICS Score	LCSC ($\|V\|$, $\|E\|$)
C. jejuni vs. *E. coli*	GHOST	24.7%	9.1%	(583, 660)
	MI-GRAAL	24.4%	8.6%	(579, 630)
	IsoRank	8.5%	0.09%	(12, 11)
A. thaliana vs. *D. mel*	GHOST	32.3%	10.0%	(1023, 1083)
	MI-GRAAL	32.2%	11.5%	(976, 1031)
	IsoRank	9.22%	0.7%	(26, 26)

Table 2: *The percentage exact overlap in* GO *annotations for aligned protein pairs. The # Terms column gives the percentage of aligned proteins that share at least the given number of GO annotations. The numbers in parentheses are P-values, measuring the statistical significance of the enrichments.*

Networks	# Terms	GHOST	MI-GRAAL	IsoRank
C. jejuni vs. *E. coli*	≥ 1	37.3 ($4.7e^{-21}$)	34.3 ($9.8e^{-13}$)	44.1 ($4.9e^{-49}$)
	≥ 2	20.6 ($3.7e^{-37}$)	14.5 ($8.6e^{-12}$)	32.9 ($5.5e^{-123}$)
	≥ 3	13.5 ($1.5e^{-45}$)	7.6 ($1.3e^{-11}$)	28.6 ($4.5e^{-191}$)
	≥ 4	10.0 ($2.0e^{-60}$)	4.0 ($4.5e^{-11}$)	20.0 ($3.3e^{-243}$)
A. thaliana vs. *D. mel*	≥ 1	61.2 ($3.2e^{-22}$)	60.1 ($4.2e^{-18}$)	63.8 ($1.7e^{-36}$)
	≥ 2	30.5 ($1.0e^{-72}$)	27.5 ($9.3e^{-50}$)	36.6($2.2e^{-137}$)
	≥ 3	19.1 ($1.1e^{-94}$)	16.9 ($3.7e^{-70}$)	25.8 ($2.2e^{-195}$)
	≥ 4	15.2 ($7.9e^{-123}$)	11.8 ($9.1e^{-73}$)	20.2 ($2.1e^{-217}$)

Table 3: *The similarities between the annotations of aligned proteins under different aspects of the* GO *hierarchy. Unlike the "exact overlap" metric used in Table 2, this similarity measure accounts for the ontological similarity and precision of the potentially different terms with which mapped proteins are labeled. The numbers in parentheses, when given, are estimated P-values. When no value is given, the result was below the approximation threshold. The similarities for the biological process (BP) and molecular function (MF) aspects were computed using the Resnik measure. However, the Resnik measure exhibited certain degeneracy issues under the cellular component (CC) aspect, where annotations for the proteins in the alignments were particularly sparse. Under this aspect, the Lin measure, which showed superior discriminative ability, was used.*

Networks	Method	s_{BP}	s_{MF}	s_{CC}
C. jejuni vs. *E. coli*	GHOST	1.75	1.36	4.0 (0.33)
	MI-GRAAL	1.27 ($1.3e^{-14}$)	1.01	4.0 (0.33)
	IsoRank	3.02	2.32	4.04 (0.15)
A. thaliana vs. *D. mel*	GHOST	2.37	1.69	2.32 ($1e^{-4}$)
	MI-GRAAL	2.23	1.59	2.27 (0.06)
	IsoRank	2.63	1.97	2.33 ($2e^{-5}$)