Græmlin: General and robust alignment of multiple large interaction networks

Jason Flannick, Antal Novak, Balaji S. Srinivasan, Harley H. McAdams and Serafim Batzoglou

Genome Res. published online Aug 9, 2006;
Access the most recent version at doi:10.1101/gr.5235706

Græmlin: General and robust alignment of multiple large interaction networks

Jason Flannick,[1,4] Antal Novak,[1,4] Balaji S. Srinivasan,[2,3] Harley H. McAdams,[2] and Serafim Batzoglou[1,5]

[1]*Department of Computer Science, Stanford University, Stanford, California 94305, USA;* [2]*Department of Developmental Biology, Stanford University, Stanford, California 94305, USA;* [3]*Department of Electrical Engineering, Stanford University, Stanford, California 94305, USA*

The recent proliferation of protein interaction networks has motivated research into network alignment: the cross-species comparison of conserved functional modules. Previous studies have laid the foundations for such comparisons and demonstrated their power on a select set of sparse interaction networks. Recently, however, new computational techniques have produced hundreds of predicted interaction networks with interconnection densities that push existing alignment algorithms to their limits. To find conserved functional modules in these new networks, we have developed Græmlin, the first algorithm capable of scalable multiple network alignment. Græmlin's explicit model of functional evolution allows both the generalization of existing alignment scoring schemes and the location of conserved network topologies other than protein complexes and metabolic pathways. To assess Græmlin's performance, we have developed the first quantitative benchmarks for network alignment, which allow comparisons of algorithms in terms of their ability to recapitulate the KEGG database of conserved functional modules. We find that Græmlin achieves substantial scalability gains over previous methods while improving sensitivity.

[Supplemental material is available online at www.genome.org. Græmlin is available at http://graemlin.stanford.edu, and source code is available under the GNU Public License.]

The publication of the second complete genome sequence in 1995 (Kaneko et al. 1995) ushered in the era of computational comparative genomics. The years that followed saw the application of cross-species genomic comparisons to problems ranging from gene prediction (Bafna and Huson 2000; Batzoglou et al. 2000; Korf et al. 2001; Alexandersson et al. 2003) to functional genomics (Pellegrini et al. 1999) to the analysis of entire genomes (Waterston et al. 2002; Cooper et al. 2004; Hillier et al. 2004). These diverse application areas were united by perhaps the most important premise of modern biology: the principle that evolutionary conservation implies functional relevance (Bejerano et al. 2004; Cooper et al. 2005; Siepel et al. 2005).

Today, the most direct analog of the exponential growth in sequence data is the rise of large-scale protein interaction network data (Uetz et al. 2000; Giot et al. 2003; Li et al. 2004). Computational and experimental techniques for inferring these networks have steadily improved (Fromont-Racine et al. 1997; Eisen et al. 1998), and state-of-the-art methods use multiple data sources to produce a unified prediction of protein interactions (Lee et al. 2004; Lu et al. 2005). The number of interaction networks is likewise increasing rapidly; in particular, a recent technique for computationally scalable data integration has produced integrated protein interaction networks for 11 microbes (Srinivasan et al. 2006), with hundreds more in preparation. Just as the rapid deposition of genomic data enabled the study of sequence conservation, the growth in network quality and availability allows us to ask questions at the network level (Milo et al. 2002).

[4]These authors contributed equally to this work.
[5]Corresponding author.
E-mail serafim@cs.stanford.edu; fax (650) 725-1449.
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.5235706. Freely available online through the *Genome Research* Open Access option.

One promising way of answering such questions is through network alignment, a systems-biological analog of sequence alignment intended to identify conserved functional modules (Hartwell et al. 1999). Research in this area has steadily progressed, beginning with manual alignments of metabolic pathways (Dandekar et al. 1999; Forst and Schulten 2001), proceeding to precursors of network alignment guided by highest-scoring pairwise BLAST hits (Altschul et al. 1997; Ogata et al. 2000; Matthews et al. 2001; Stuart et al. 2003), and culminating in the modern formulation of network alignment (Kelley et al. 2003; Koyuturk et al. 2005). Recent work has partially removed previous limitations by enabling searches for conserved multiprotein complexes in addition to pathways (Sharan et al. 2005a) and allowing the simultaneous comparison of three species rather than two (Sharan et al. 2005b). However, the general problem of finding conserved modules of arbitrary topology within an arbitrary number of networks has not yet been addressed.

In this paper we describe Græmlin, a novel network alignment framework that is fast, scalable, and capable of searching large sets of dense networks for conserved functional modules. Græmlin's probabilistic formulation of the topology-matching problem eliminates earlier restrictions on the possible architecture of conserved modules. Most important, Græmlin is the first program capable of multiple alignment of an arbitrary number of networks.

To assess Græmlin's ability to find conserved functional modules, we have performed the first quantitative comparison of network aligners. Using data sets containing known biological modules as a benchmark (Ashburner et al. 2000; Kanehisa and Goto 2000), we find that Græmlin achieves substantial gains in sensitivity over previous methods while offering fast and scalable searches of multiple, large networks. In addition to statistical

benchmarking, we present detailed analyses of several alignments that suggest interesting hypotheses about protein function.

Græmlin is available through a Web interface located at http://graemlin.stanford.edu, where users can search for conserved functional modules within a large database of microbial networks. Source code is also available under the GNU Public License.

Methods

Græmlin is a network alignment algorithm capable of searching large sets of dense interaction networks for evolutionarily conserved functional modules, which are groups of homologous proteins with conserved pairwise interactions. Græmlin supports both global and local search; it can be used either to generate an exhaustive list of conserved modules in a set of networks (network-to-network alignment) or to find matches to a particular module within a database of interaction networks (query-to-network alignment).

Depending on the context of a study, one may wish to find functional modules that are present within all species or simply those that are enriched within a particular clade. Græmlin enables both kinds of comparative analysis, as it can rapidly search a large number ($N > 3$) of interaction networks to find functional modules that are significantly conserved in two or more species.

The efficient performance of Græmlin is due to the use of several strategies common in sequence alignment (Batzoglou 2005). First, its variant of "progressive alignment" (Feng and Doolittle 1987) allows it to scale linearly with the number of networks compared. Second, Græmlin searches for pairwise alignments between networks using a modification of the "seed extension" method popularized by BLAST (Altschul et al. 1997). Finally, it allows an explicit speed-sensitivity trade-off through the control of a parameter analogous to the BLAST word size (Altschul et al. 1990).

Below we outline our definition of a network alignment, the scoring model used by Græmlin, and its algorithm for finding high-scoring pairwise and multiple alignments.

Definition of an alignment

Given interaction networks for a set of related species, the goal of a network aligner is to extract conserved subnetworks that are statistically significant relative to alignments found in biologically unrelated networks. Such subnetworks are hypothesized to have evolved from a functional module originally present in the common ancestor.

We represent each input network as a weighted graph $G_i = (V_i, E_i)$, where nodes correspond to proteins and each weighted edge specifies the probability that two proteins interact. We define a network alignment as a set of subgraphs chosen from the interaction networks of different species, together with a mapping between corresponding, or aligned, proteins. To uniquely specify an alignment, we require that the mapping be transitive; that is, if protein A is aligned to proteins B and C, then protein B must also be aligned to protein C. Mathematically, this means that the mapping is an equivalence relation; consequently, the groups of aligned proteins are disjoint, and we refer to them as equivalence classes for this reason.

We also require that all aligned proteins be homologous. Therefore, proteins in the same equivalence class are in general members of the same protein family (Andreeva et al. 2004; Marchler-Bauer et al. 2005). In this manner, a biological interpretation of an alignment is a collection of protein families whose interactions are conserved across a given set of species.

This definition affords important advantages. Because the members of a protein family descend from a common ancestor, we can reconstruct the evolutionary events leading from each ancestral protein to its extant descendants. By combining this with a reconstruction of the evolutionary history of each pairwise interaction, we can interpret each network alignment as a hypothesis about the evolution of a conserved ancestral module. Intuitively, network alignments should receive high scores if their evolutionary dynamics resemble those of known, conserved functional modules rather than those of random collections of proteins.

With this definition, there are two core problems in network alignment. First, we must devise a scoring framework that captures the knowledge we have about module evolution. Then, we must find a way to rapidly identify high-scoring alignments—meaning conserved functional modules—from among the exponentially large set of possible alignments. We address each problem in turn.

Scoring of an alignment

The evolutionary interpretation of an alignment leads to a natural scoring function. We first define two models that assign probabilities to the evolutionary events leading from the hypothesized ancestral module to modules in the extant species: the alignment model $\mathcal{M}$ posits that the module is subject to evolutionary constraint, while the random model $\mathcal{R}$ assumes that the proteins are under no constraints. The score of the alignment is the log-ratio of the two probabilities, a common method for scoring sequence alignments (Durbin et al. 1998). Figure 1 shows a sample alignment, together with an overview of the scoring framework.

Græmlin individually scores each equivalence class and each edge of an alignment. To score equivalence classes, it uses a straightforward scheme that reconstructs the most parsimonious ancestral history of an equivalence class, based on five types of evolutionary events: protein sequence mutations, protein insertions and deletions, protein duplications, and protein divergences; a protein divergence occurs when a paralogous protein loses its function and is the inverse of a duplication. The models $\mathcal{M}$ and $\mathcal{R}$ give each of these events a different probability. Currently, we estimate probabilities of sequence mutations in a principled manner, but we determine probabilities of other events heuristically; a detailed discussion is provided in the Supplemental material. This is analogous to sequence alignment, where traditionally substitution matrices are estimated rigorously (Henikoff and Henikoff 1993; Chiaromonte et al. 2002) but gap penalties are set in a heuristic manner (Brudno et al. 2003; Blanchette et al. 2004; Bray and Pachter 2004; Edgar 2004).

To determine the probabilities for sequence mutations, Græmlin uses weighted sum-of-pairs scoring (Altschul et al. 1989). Each model assigns a probability to a pair of proteins based on a BLAST bitscore; we trained the alignment model $\mathcal{M}$ by sampling pairs of proteins from within the same COG (Tatusov et al. 1997; Kelley et al. 2003; Sharan et al. 2005b), and we trained the random model $\mathcal{R}$ on random pairs of proteins. The log-ratio of these two distributions gives a scoring function for a pair of proteins: The sequence mutation score of an equivalence class is the weighted sum-of-pairs score taken over all pairs of proteins in the class, using a phylogenetic tree relating the species in the alignment.

As with equivalence classes, we define edge scores as the log-ratio of two probabilities: Each edge e is assigned a score $S_e = \log(\Pr_{\mathcal{M}}(e)/\Pr_{\mathcal{R}}(e))$.

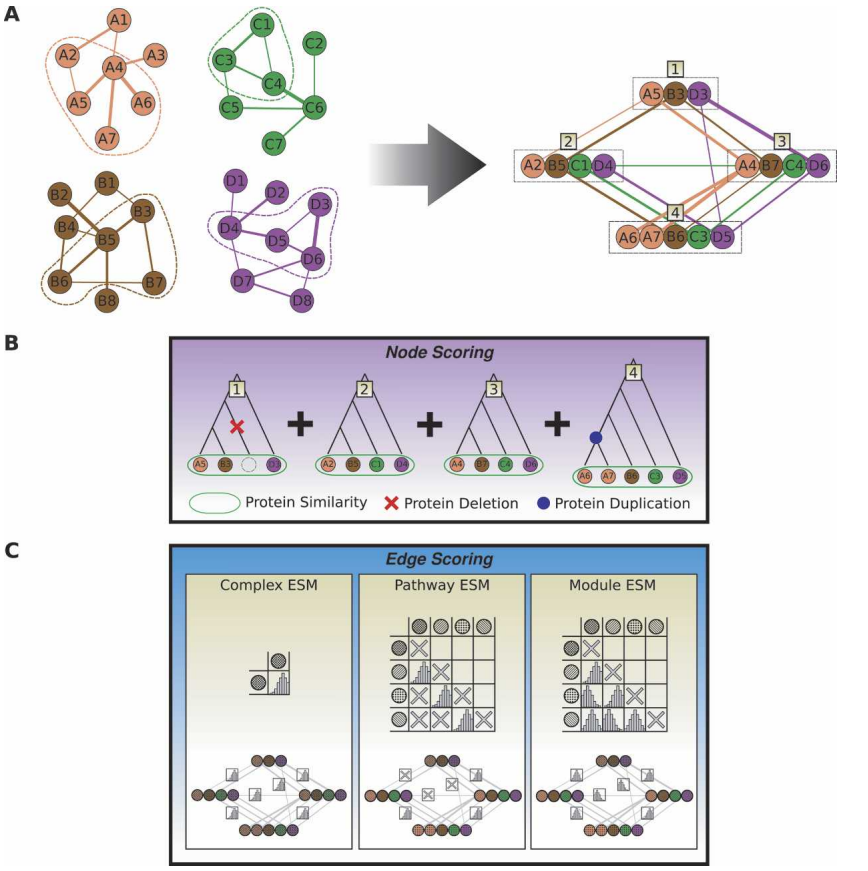

Figure 1. Method for scoring a multiple network alignment. (*A*) A sample multiple alignment. The four networks are from four different species. Each circle represents a protein, and edges link proteins that are hypothesized to interact; the width of an edge is proportional to the probability that an interaction is present. This sample alignment of the networks consists of four equivalence classes, numbered 1 through 4. (*B*) Node scoring method. Græmlin first scores each equivalence class independently by reconstructing the most parsimonious ancestral history of the involved proteins and then assessing penalties for sequence mutations and protein insertions, deletions, duplications, and divergences. Græmlin scores sequence mutations by using weighted sum-of-pairs scoring, obtaining pairwise scores based on BLAST results of the proteins; it scores all other events using heuristic constants. (*C*) Edge scoring method. Græmlin scores each edge using an Edge Scoring Matrix (ESM) as described in the text. For illustrative purposes, three alternative ESMs are shown, together with how Græmlin would score the alignment using each of them. The first ESM rewards protein complexes by specifying that edges between any pair of equivalence classes should have high weight; the matrix has only one cell because every edge is scored with the same distribution. The Complex ESM will score the alignment fairly highly but will penalize it because of the missing edges between equivalence classes 1 and 4 as well as 2 and 3. The Pathway ESM assigns a higher score to the alignment because it does not require high weight edges between all pairs of equivalence classes. It achieves this by a four-row matrix, where each label corresponds to a distinct node in a four-protein pathway. Edges between adjacent nodes in the pathway have high weights, and all other edges can have high or low weights without affecting the score; "don't care" distributions, symbolized by an X in the matrix, assign a score of 0 to those edges. The Module ESM assigns an even higher score to the alignment by conforming exactly to its structure; such an ESM is useful when a known module in one species is used as a query for searching another network.

The random model $\mathcal{R}$ assigns each edge a probability parametrized not only by its weight but also by the degrees of its endpoints; this captures the intuitive notion that in any graph, two nodes of high degree are more likely to interact by chance than two nodes of low degree. The alignment model for edges, however, is not as straightforward: Unlike in the case of proteins, Græmlin cannot always assume that an edge existed in the ancestral module. This assumption would, for instance, always reward highly connected modules more than equally conserved but loosely connected modules. The alternative of considering only the conservation of the presence or absence of an edge would score a completely unconnected alignment highly.

To address these issues, Græmlin uses a novel scoring scheme that allows a user to specify the desired ancestral topology; this generalizes previous edge-scoring approaches (Sharan et al. 2005b) and permits searches for arbitrary module structures, including as special cases multiprotein complexes and pathways. We use an Edge Scoring Matrix, or ESM, to encapsulate the desired module structure into a symmetric matrix. An ESM has a set of labels by which its rows and columns are indexed, and each cell in the matrix contains a probability distribution over edge weights. To score edges in an alignment, Græmlin first assigns to each equivalence class one of the labels from the ESM. Then, it scores each edge e using the cell in the matrix indexed by the labels of the two equivalence classes to which its endpoints belong: The function in the cell maps the weight of the edge to a probability $\Pr_{\mathcal{M}}(e)$, which is used to compute the score $S_e = \log(\Pr_{\mathcal{M}}(e)/\Pr_{\mathcal{R}}(e))$.

Figure 1C shows three examples of an ESM, as well as the functions used by each to score the edges in a sample alignment; these include two special cases, pathways and multiprotein complexes, that have been the subjects of past studies (Kelley et al. 2003; Koyuturk et al. 2005; Sharan et al. 2005b). To search for conserved multiprotein complexes, we use a Complex ESM, which consists of a single label with an alignment distribution assigning high probabilities to high edge weights. A Pathway ESM has one label for each protein in the pathway and rewards high edge weights between adjacent proteins; between all other proteins, the alignment and random distributions are the same, so that Græmlin neither rewards nor penalizes edges connecting nonadjacent proteins.

The third type of ESM we consider is automatically generated when a user searches a large network for matches to a small query network. We refer to this as a Module ESM because the query network will often consist of a hypothetical or known biological module. In this case, Græmlin creates a label in the ESM for each node in the query and generates the alignment distribution based on the edges that are present or absent in the query. For each cell in the ESM, it defines the distribution based on the weight of the edge between the two corresponding proteins in the query. When aligning a query to multiple species, Græmlin refines the ESM as more species are added to the alignment; in this case, rather than creating a label for each protein, it creates a label for each equivalence class and uses kernel density estimation (Duda et al. 2000) to train the distributions from the entire set of edges present in the alignment. When used in this form, we refer to an ESM as a

Location Specific Scoring Matrix, or LSSM, because of its conceptual and practical similarity to the Position Specific Scoring Matrix (PSSM) used in PSI-BLAST (Altschul et al. 1997). Figure 2 shows a simple example of the manner in which Græmlin constructs an LSSM.

Alignment algorithm

Figure 3 shows an outline of the Græmlin algorithm, including the methodology it uses for pairwise and multiple alignment.

Pairwise alignment

To search for high-scoring alignments between a pair of networks efficiently, Græmlin first generates a set of "seeds," which it uses to restrict the size of the search space. We refer to the structures used for seed generation as "*d*-clusters," which consist of *d* proteins that are close together in a network and are analogs of *k*-mers in seeded local alignment search.

For each network, Græmlin constructs one *d*-cluster for every node by finding the $d - 1$ nearest neighbors of that node, where the length of an edge is the negative logarithm of its weight. Græmlin compares two *d*-clusters D_1 and D_2 by mapping a subset of nodes in D_1 to a subset of nodes in D_2 and reporting a score equal to the sum of all pairwise scores induced by the mapping; the score of two *d*-clusters is the highest-scoring such mapping. Græmlin identifies pairs of *d*-clusters, one from each network, that score higher than a threshold *T* and uses these as seeds. Figure 3B shows a sample set of *d*-clusters generated from two networks, as well as a high-scoring pair.

The benefits of the *d*-cluster seeding technique are severalfold. First, Græmlin can compare *d*-clusters rapidly, since the comparison neglects edge scores. Second, the parameters *d* and *T* allow for a speed-sensitivity trade-off. As an example, a lower value of *T* will achieve higher sensitivity but require increased running time; this adjustable trade-off is not present in previous techniques (Koyuturk et al. 2005; Sharan et al. 2005b). Finally, high-scoring alignments are likely to contain high-scoring *d*-clusters, since a high node score of an alignment is usually a prerequisite to a high overall score. We can give this intuition a mathematical foundation using ideas similar to those underlying spaced seed analysis techniques (Ma et al. 2002; Sun and Buhler 2005); this analysis, which we discuss in the Supplemental material, yields some intuition into the interplay between the two dependent parameters *d* and *T*.

Given two networks, Græmlin enumerates the set of seeds between them and tries to transform each, in turn, into a high-scoring alignment. In a manner similar to that used in existing methods (Koyuturk et al. 2005; Sharan et al. 2005b), the seed extension phase is greedy and occurs in successive rounds. At each step, all proteins adjacent to some node in the alignment constitute the "frontier," which contains candidates to be added to the alignment. Græmlin selects from the frontier the pair of proteins that, when added to the alignment, yields the maximal increase in score; the extension phase stops when no pair of proteins on the frontier can increase the score of the alignment. Figure 3C illustrates the extension algorithm. Græmlin uses several heuristics to control for the exponential increase in the size of the frontier as it adds more nodes to the alignment.

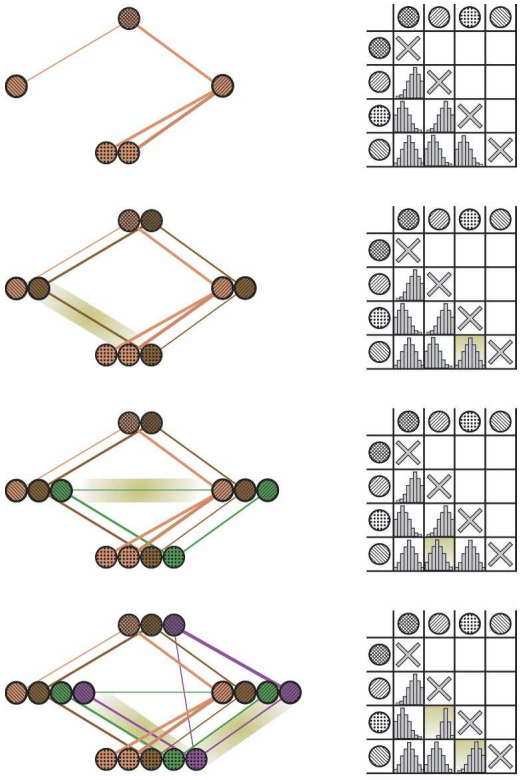

Figure 2. An example of an LSSM. As Græmlin successively adds species to the multiple alignment, the distributions in the ESM cells change to reflect the new edges. At each step, the cell with a modified distribution is highlighted together with the edge that caused the change.

Multiple alignment

Græmlin performs multiple alignment using an analog of the progressive alignment technique commonly used in sequence alignment. Using a phylogenetic tree, it successively aligns the closest pair of networks, constructing several new networks from the resulting alignments. Græmlin places each new network at the parent of the pair of networks that it just aligned. The constructed networks contain nodes that are no longer proteins but equivalence classes, but all scoring and alignment methodologies readily generalize to such networks. Græmlin continues this process until the only remaining networks are at the root of the phylogenetic tree.

To enable comparisons of unaligned parts of a network to more distant species as it traverses the phylogenetic tree, rather than construct a network only from the high-scoring alignments, Græmlin also maintains two additional networks composed of the unaligned nodes from the two original networks. For example, in Figure 3D, Græmlin constructs three networks from the original two that it aligns; as a result, in Figure 3E, the parent of these two networks contains one network for each possible subset of its children. The end result is that after completion of the entire multiple alignment, Græmlin produces multiple alignments of all possible subsets of species. It avoids an exponential running time in practice because after each pairwise alignment, the networks it constructs have small overlaps. The total number of nodes in all networks therefore does not increase significantly.

Results

We measured the performance of Græmlin by assessing its ability to align known biologically functional modules. We compared it to two alignment algorithms, NetworkBLAST (Sharan et al. 2005b) and MaWISh (Koyuturk et al. 2005); because the focus of MetaPathwayHunter (Pinter et al. 2005) is different from general network alignment, we did not include it in our tests. We tested these methods on a set of 10 microbial protein interaction net-

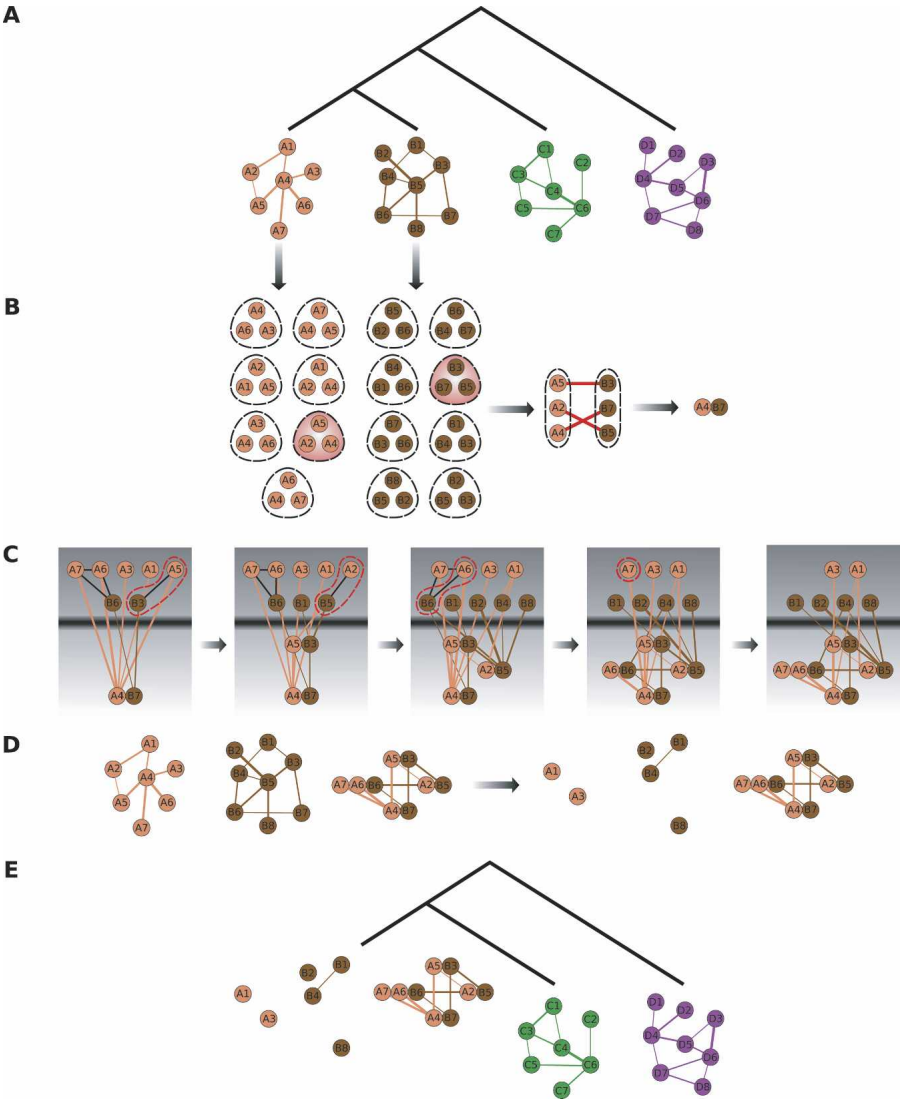

Figure 3. Outline of the Græmlin algorithm. (*A*) Shown here are four networks, together with their phylogenetic relationship. Græmlin will multiply align all four. (*B*) Græmlin first performs a pairwise alignment of the two closest species. It generates a set of *d*-clusters from each network; each node and its $d - 1$ closest neighbors constitute a *d*-cluster. Græmlin scores all pairs of *d*-clusters by finding for each pair the highest scoring mapping among nodes and selects the pairs that score greater than a user-specified threshold *T*; one particular high-scoring pair is highlighted together with the node mapping responsible for its score. Græmlin transforms all high-scoring pairs into seeds by aligning the two highest scoring nodes. (*C*) Græmlin extends the seed using a greedy algorithm; each extension step is shown in a gray box. At each step, Græmlin adds to the frontier all nodes connected to a node currently in the alignment; the frontier is shown in the *upper* half of each box. Græmlin adds to the alignment the pair of nodes or single node from the frontier that maximally increases the alignment score; the extension phase stops when no nodes from the frontier can augment the alignment score. (*D*) Græmlin transforms the resulting alignment, together with the unaligned nodes from the two original networks, into three generalized networks for use in the next phase of progressive alignment. (*E*) In the next step of progressive alignment, Græmlin will perform three pairwise alignments, one for each of the newly created generalized networks. Its running time will not scale by a factor of 3, however, as the total number of nodes in all networks remains roughly the same.

works constructed via the SRINI algorithm (Srinivasan et al. 2006), which generates weighted interaction networks by integrating a set of functional predictors, such as coexpression, coinheritance, coevolution, and colocation, and computing the interaction probability for each pair of proteins. Details on the methodology for constructing these networks are included in the Supplemental material.

We assessed the sensitivity of each method by counting the number of KEGG pathways that it aligned between two species (Kanehisa and Goto 2000). We identified a KEGG pathway as "hit" if a method aligned at least three proteins in the pathway to their counterparts in the other species. We defined the "coverage" of a pathway to be the fraction of proteins correctly aligned within the pathway. Changing the definition of a hit pathway to require two or four, instead of three, proteins did not affect the relative performance of the aligners.

We did not use all KEGG pathways for these comparisons, as SRINI does not accurately recapitulate each one. We therefore curated the set of KEGG pathways by ignoring all that did not have a connected component of size at least three in each of the assessed networks. To avoid biasing results toward a specific algorithm, we did not further curate the set by examining the conservation of the pathways. We refer to each pathway in the curated set as an "alignable" KEGG pathway.

As one measure of specificity, we computed the number of "enriched" alignments. To calculate enrichment, we first assigned to each protein all of its annotations from level eight or deeper in the GO hierarchy (Ashburner et al. 2000); given an alignment, we then discarded unannotated proteins and calculated its enrichment using the GO TermFinder (Boyle et al. 2004). We considered an alignment to be enriched if the *P*-value of its enrichment was <0.01.

As a second measure of specificity, we counted the fraction of nodes that have KEGG orthologs but were aligned to any nodes other than their KEGG orthologs. Both this measure and calculations of enrichment are imperfect measures of specificity, but they work as rough guides to ensure that an aligner is not completely sacrificing specificity to increase sensitivity.

We also assessed multiple alignment algorithms using these metrics. When evaluating the sensitivity metric, we identified a KEGG pathway as hit if a method aligned at least three proteins in each species to their correct counterparts in all other species. We measured specificity by computing enrichments and counting misaligned nodes exactly as in the case of pairwise alignments.

To our knowledge, the only other quantitative tests of alignment quality measured the accuracy of predictions of interactions and protein function in eukaryotic networks (Kelley et al. 2003; Sharan et al. 2005b). The first such test is relevant mainly

Table 1. Network statistics

Species	Color	# Nodes	Edge Threshold	# Edges	# Edges per Node	# Alignable KEGGs
Campylobacter jejuni NCTC 11168		1629	0.25	22116	13.58	42
			0.5	6171	3.79	29
Caulobacter crescentus		3737	0.25	40568	10.86	70
			0.5	6018	1.61	55
Escherichia coli K12		4242	0.25	216426	51.02	72
			0.5	35132	8.28	70
Helicobacter pylori 26695		1576	0.25	12960	8.22	32
			0.5	3723	2.36	26
Mycoplasma tuberculosis H37Rv		3991	0.25	129183	32.37	75
			0.5	17380	4.35	61
Salmonella typhimurium LT2		4527	0.25	94609	20.90	61
			0.5	18149	4.01	55
Streptococcus pneumoniae TIGR4		2094	0.25	25732	12.29	29
			0.5	4607	2.20	23
Streptomyces coelicolor		8154	0.25	230467	28.26	76
			0.5	60852	7.46	54
Synechocystis PCC 6803		3166	0.25	69439	21.93	47
			0.5	13963	4.41	32
Vibrio cholerae		3835	0.25	36087	9.41	61
			0.5	7886	2.06	45
Saccharomyces cerevisiae	N/A	4766	N/A	15200	3.19	22
Caenorhabditis elegans	N/A	2629	N/A	3950	1.50	0
Drosophila melanogaster	N/A	7067	N/A	21822	3.09	4

This table shows various statistics about the interaction networks of the species on which we ran tests, as well as three eukaryotic networks for comparative purposes. For each species, we tested two interaction networks: one with edge weights below 0.25 removed and one with edge weights below 0.5 removed. For the networks of each species, the columns present the total number of nodes, the total number of edges, the average number of edges per node, and the number of KEGG pathways containing a connected component of size at least three. The table also shows the color that represents each species; these colors correspond to those used to identify species in the figures presenting alignments.

in networks with a high number of false positives in one particular species; this is not the case in the microbial networks on which we tested, as three functional predictors used for their construction incorporate some measure of cross-species conservation. As the second test overlaps considerably with our tests measuring enrichment, we do not present results beyond those measuring our notions of sensitivity and specificity.

One issue with the networks constructed by SRINI is that they are complete; that is, SRINI assigns an interaction probability to every pair of proteins. Because these networks are intractably large for any existing algorithm, we thresholded them by removing low-weight edges before running our tests. We generated two sets of networks: one with an edge threshold of $P \geq 0.25$ and another with a threshold of $P \geq 0.5$. Table 1 lists the species on which we ran tests, in addition to statistics on the network sizes and presence of KEGG pathways in the networks. For comparison purposes, the table also shows the same statistics for the eukaryotic networks that previous studies on alignment have analyzed (Xenarios et al. 2002; FlyBase Consortium 2003; Christie et al. 2004; Harris et al. 2004). Table 2 lists, for each subset of species on which we tested, the number of alignable KEGG pathways present in all species in the subset.

We did not test on the eukaryotic networks because our sensitivity metric is inapplicable on them; as Table 1 shows, they recapitulate essentially no KEGG pathways. While in principle one could define other sensitivity metrics, the greater quality of the SRINI networks provides a much simpler and straightforward test scenario. In addition, the SRINI networks are much larger than the eukaryotic networks and consequently provide a better test of the scalability of an algorithm.

For all tests and all alignment algorithms, we considered alignments with *P*-values <0.01 as high-scoring; for each test case, we calculated *P*-values by sampling from a large number of runs on random data sets. We constructed the random data sets

Table 2. KEGG pathway conservation statistics

Species set	Threshold	No. of alignable KEGGs
E. coli, C. crescentus	0.25	55
	0.5	44
E. coli, M. tuberculosis	0.25	60
	0.5	54
E. coli, V. cholerae	0.25	54
	0.5	39
E. coli, S. coelicolor	0.25	57
	0.5	43
E. coli, C. crescentus, V. cholerae	0.25	47
	0.5	27
C. jejuni, E. coli, H. pylori	0.25	28
	0.5	15

This table shows the number of alignable KEGG pathways that are present for various subsets of species. An alignable KEGG pathway is present for a given subset of species if the pathway is alignable in each of the species in the subset.

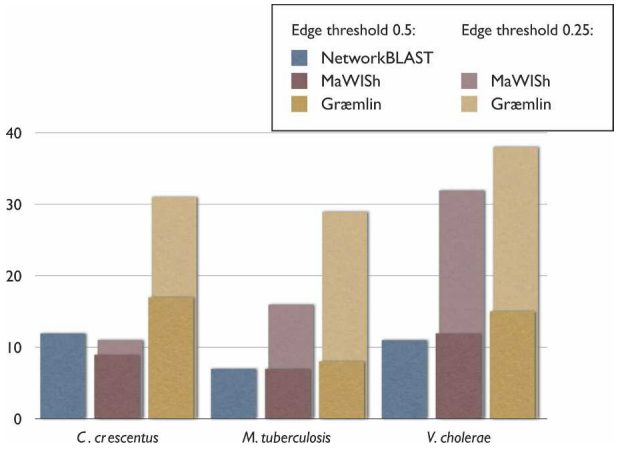

Figure 4. Sensitivity comparison of methods. For three pairwise alignments of *E. coli*, shown are the number of KEGGs hit by each aligner. For Græmlin and MaWISh, this graph includes results on networks with edge thresholds of both hold and 0.5. For NetworkBLAST, however, we only include results on networks thresholded at 0.5, as it did not scale to denser inputs.

using techniques similar to those used in previous approaches by redistributing the edges of a real network while maintaining the original node degree distribution (Kelley et al. 2003; Koyuturk et al. 2005; Pinter et al. 2005; Sharan et al. 2005a). Unless noted otherwise, we ran all aligners with their default parameters. We performed all tests on a 2.8 GHz Intel Xeon processor with 2 Gb of RAM running the Linux operating system.

Network-to-network alignment

The goal of complete network-to-network alignment is to find conserved subcomponents of networks, and the results often suggest potential functional modules present in more than one species. Our first set of tests assessed performance of each algorithm under this application; this is the focus of both MaWISh, which searches for conserved heavy subgraphs, and NetworkBLAST, which searches for conserved protein complexes and pathways.

With all three methods, we aligned the networks of *Escherichia coli* K12 and *Caulobacter crescentus*; *E. coli* and *Mycoplasma tuberculosis* H37Rv; *E. coli* and *Vibrio cholerae*; and *E. coli* and *Streptomyces coelicolor*. Owing to its excessive running time on networks with the lower edge threshold, we report results for NetworkBLAST only on networks with a threshold of 0.5. In addition, MaWISh cannot perform alignments on *S. coelicolor* because for each input species it requires COG data (Tatusov et al. 1997), which is not available for *S. coelicolor*. Figure 4 summarizes sensitivity for three of the test cases on networks with edge thresholds of 0.25 and 0.5, and Table 3 shows more detailed results on networks thresholded at 0.5. Complete results are included in the Supplemental material.

These results show that there is a legitimate reason for using networks with a lower edge threshold, since the sensitivities of both MaWISh and Græmlin drop dramatically on networks with a higher threshold without a corresponding increase in specificity. Consequently, the ability of both methods to scale to data sets of this size is important.

When searching for highly connected components, or multiprotein complexes, Græmlin is significantly more sensitive than the other two methods, both with respect to the number of KEGGs hit and with respect to the average coverage of a KEGG, without sacrificing specificity. It also aligns significantly more nodes overall than the other two methods without misaligning a higher number of proteins.

When searching for pathways, Græmlin is more sensitive than NetworkBLAST, although both methods are not as sensitive as those that search for multiprotein complexes. This is predominantly because a pathway alignment must be much larger than an alignment of multiprotein complexes to be statistically significant. For example, if an alignable KEGG pathway contains a clique of four proteins, it will score highly as both a multiprotein complex and a pathway. However, because four-node cliques are much less likely to occur in unrelated networks than four-node pathways, the multiprotein complex alignment will be more statistically significant. As most alignable KEGGs appear as cliques in the SRINI microbial networks, searches for highly connected components are consequently more sensitive than pathway searches with respect to the metric of hitting KEGG pathways. However, past studies have shown that pathway searches do have uses beyond identifying conserved modules (Kelley et al. 2003; Sharan et al. 2005b).

Table 3. Results on pairwise alignment of complete networks thresholded at 0.5

		KEGGs hit	KEGG coverage	Alignments enriched	Running time (sec)
E. coli vs. *C. crescentus*					
MaWISh		9 (20%)	32%	72%	3
NetworkBLAST	Pathway	6 (14%)	28%	61%	9624
	Complex	12 (27%)	49%	72%	
Græmlin	Pathway	15 (34%)	47%	68%	21
	Complex	17 (39%)	45%	67%	11
E. coli vs. *M. tuberculosis*					
MaWISh		7 (13%)	20%	85%	3
NetworkBLAST	Pathway	7 (13%)	24%	88%	301
	Complex	7 (13%)	32%	88%	
Græmlin	Pathway	8 (15%)	36%	89%	11
	Complex	8 (15%)	39%	89%	8
E. coli vs. *V. cholerae*					
MaWISh		12 (31%)	35%	64%	3
NetworkBLAST	Pathway	10 (26%)	35%	58%	8797
	Complex	11 (28%)	41%	64%	
Græmlin	Pathway	19 (49%)	48%	75%	13
	Complex	15 (38%)	55%	74%	12
E. coli vs. *S. coelicolor*					
MaWISh		N/A	N/A	N/A	N/A
NetworkBLAST	Pathway	6 (14%)	23%	46%	122,168
	Complex	10 (23%)	67%	95%	
Græmlin	Pathway	8 (19%)	58%	88%	734
	Complex	9 (21%)	59%	85%	829

For each pair of species, we performed complete network-to-network alignment using MaWISh and Græmlin. For each tested method, shown, from left, is the total number of KEGG pathways hit by an alignment, the fraction of KEGG pathways hit by an alignment, the average coverage of a KEGG pathway, the percentage of enriched alignments, and the total running time. We calculated the average coverage of KEGGs with respect to only those KEGGs that an aligner hit, and measured running time in CPU-seconds.

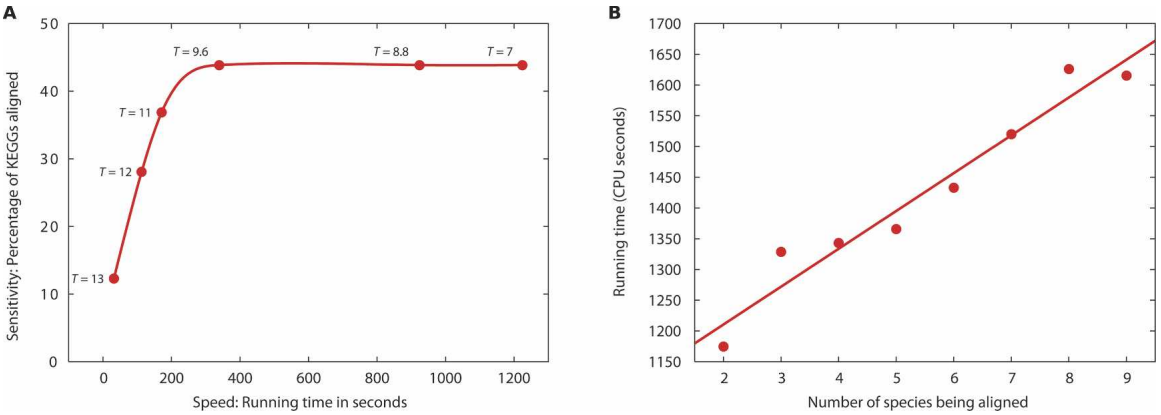

Figure 5. Running-time performance of Græmlin. (*A*) The speed sensitivity trade-off. Each point represents a run of Græmlin with $d = 4$ and different values of T. For each set of parameters, the *x*-axis plots the running time, and the *y*-axis plots the fraction of alignable KEGGs hit. (*B*) Progressive multiple alignment. Beginning with *E. coli*, we added species of increasing evolutionary distance to the multiple alignment. The pairwise running time is comparatively high because the two species aligned, *E. coli* and *S. typhimurium*, are the two most similar species and have many high-scoring alignments. In this manner, adding particularly close species to the alignment can lead to higher-than-average increases in running time, but over all species the average scaling will remain roughly linear.

With respect to running time, only MaWISh and Græmlin can efficiently search the large networks on which we tested. While MaWISh is the faster of the two methods, the running time of Græmlin is comparable.

Of all the test cases, Græmlin and NetworkBLAST take the longest to run on *E. coli* versus *S. coelicolor*, primarily because of the size of the *S. coelicolor* network and the large number of homologs between these species. In this case, Græmlin can sacrifice sensitivity for speed by adjusting the parameters it uses for *d*-clusters. Figure 5A demonstrates the impact of T on running time and sensitivity. Running with its default parameters ($d = 4$, $T = 7$) on networks thresholded at 0.25, it finds 25 KEGG pathways in 1224 sec, but a slight increase in T yields 21 KEGGs in only 339 sec.

Multiple network alignment

We also performed complete three-way alignments of (*E. coli*, *C. crescentus*, *V. cholerae*) and (*E. coli*, *Campylobacter jejuni*, *Helicobacter pylori* 26,695) using Græmlin and NetworkBLAST. Table 4 shows the results of these tests.

On the networks thresholded at 0.5, NetworkBLAST hits slightly more KEGGs than Græmlin; however, Græmlin covers a much higher fraction of each KEGG and also misaligns fewer nodes. In addition, Græmlin is orders of magnitude faster than NetworkBLAST; on one of our test cases, the latter did not complete after running for more than 2 mo. Because Græmlin scales effectively to large network sizes, it can efficiently multiply align networks with a low edge threshold. This is important because the networks with a low edge threshold contain many more conserved KEGGs than the high-thresholded networks, as evinced by the dramatically increased sensitivity of Græmlin on this data set.

To further stress the scalability of Græmlin with respect to the number of species in a multiple alignment, Figure 5B shows the running times of Græmlin as it includes more species in the alignment. The roughly linear relation of running time to the number of species demonstrates the benefit of the progressive alignment technique.

Query-to-network alignment

Query-to-network alignment is a network analog of the BLAST algorithm; the goal is to search a large database of alignments for matches to an input query that is typically a hypothetical or known functional module. Both MaWISh and NetworkBLAST can perform query-to-network alignment by treating the query as a full network. On the other hand, Græmlin supports fast alignment of many queries to the same database by building an index as a one-time expense and maintaining it in memory for many successive queries.

Table 4. Results on multiple alignment of complete networks

		KEGGs hit	KEGG coverage	Alignments enriched	Running time (sec)
0.25 threshold					
E. coli vs. *C. crescentus* vs. *V. cholerae*					
Græmlin	Pathway	27 (57%)	68%	72%	329
	Complex	29 (62%)	71%	79%	251
E. coli vs. *C. jejuni* vs. *H. pylori*					
Græmlin	Pathway	16 (57%)	57%	87%	44
	Complex	17 (61%)	63%	89%	43
0.5 threshold					
E. coli vs. *C. crescentus* vs. *V. cholerae*					
NetworkBLAST	Pathway	N/A	N/A	N/A	$>10^6$
	Complex				
Græmlin	Pathway	7 (26%)	67%	72%	63
	Complex	9 (33%)	62%	75%	38
E. coli vs. *C. jejuni* vs. *H. pylori*					
NetworkBLAST	Pathway	5 (33%)	41%	94%	32,394
	Complex	4 (27%)	38%	96%	
Græmlin	Pathway	3 (20%)	74%	82%	12
	Complex	3 (20%)	72%	79%	12

We performed three-way multiple network alignment using NetworkBLAST and Græmlin; the columns in this table are analogous to those in Table 3.

Our final set of tests assessed the performance of each method on query-to-network search; we searched *E. coli* against *C. crescentus*, *C. crescentus* against *E. coli*, *E. coli* against *M. tuberculosis*, and *M. tuberculosis* against *E. coli*. Table 5 shows the results of these tests; more detailed results are included in the Supplemental material.

For this test, sensitivity and specificity results are similar to those in the case of complete network alignment. One major difference is the relative running times of Græmlin and MaWISh; they are comparable when the database is *C. crescentus*, the smallest network, but Græmlin is much faster on the other networks. This is because Græmlin can amortize its indexing step over all queries and shows Græmlin's strength as a database search tool.

In this test case, Græmlin performs comparably when using both the Pathway ESM and the Complex ESM. The Module ESM does not offer dramatic improvements over the other two ESMs, but it does give slightly higher KEGG coverage and misalign slightly fewer nodes. This is because most KEGGs that are alignable are highly connected, making the Complex ESM close to optimal under our metrics.

Biological applications

We used Græmlin to perform a 10-way alignment of *E. coli*, *Salmonella typhimurium*, *V. cholerae*, *C. crescentus*, *C. jejuni*, *H. pylori*, *Synechocystis*, *S. coelicolor*, *M. tuberculosis*, and *S. pneumoniae*. This generated roughly 2000 significant multiple alignments, each containing all or a subset of the 10 species; complete results are available in the Supplemental material. Because the analysis of these alignments is a research direction in its own right, we selected interesting alignments manually. Our focus was predominantly on results that illustrate the utility of the various features of Græmlin.

Table 5. Results on alignment of a query network to a database thresholded at 0.5

		KEGGs hit	KEGG coverage	Running time (sec)
E. coli vs. *C. crescentus*				
MaWISh		15 (34%)	31%	37
NetworkBLAST	Pathway	8 (18%)	32%	3453
	Complex	10 (23%)	49%	
Græmlin	Pathway	20 (45%)	45%	17
	Complex	20 (45%)	47%	3
	Module	20 (45%)	48%	23
C. crescentus vs. *E. coli*				
MaWISh		9 (20%)	32%	130
NetworkBLAST	Pathway	10 (23%)	37%	4788
	Complex	10 (23%)	41%	
Græmlin	Pathway	15 (34%)	39%	6
	Complex	15 (34%)	42%	5
	Module	15 (34%)	42%	33
E. coli vs. *M. tuberculosis*				
MaWISh		10 (19%)	19%	93
NetworkBLAST	Pathway	12 (22%)	23%	3947
	Complex	12 (22%)	29%	
Græmlin	Pathway	17 (31%)	31%	3
	Complex	17 (31%)	35%	3
	Module	17 (31%)	35%	22
M. tuberculosis vs. *E. coli*				
MaWISh		6 (11%)	12%	138
NetworkBLAST	Pathway	10 (19%)	19%	5047
	Complex	7 (13%)	22%	
Græmlin	Pathway	13 (24%)	25%	5
	Complex	14 (26%)	26%	5
	Module	14 (26%)	27%	28

For each pair of species, using MaWISh, NetworkBLAST, and Græmlin, we successively aligned each KEGG pathway in the query species to the complete network of the database species. For each tested method, shown, from left, is the total number of KEGG pathways with a database hit, the fraction of KEGG pathways with a database hit, the average coverage of a KEGG pathway, and the total running time. As NetworkBLAST does not have an option to search separately for pathways and complexes, the table lists the combined running time of both searches.

Functional annotation

Network alignment can be used to assign roles to proteins of unknown function in two ways. First, "annotation transfer" assigns to a protein of unknown function the annotation of a protein to which it is aligned. This procedure is similar to the traditional method of annotation transfer using only sequence alignment, but network alignment strengthens the hypothesis by revealing conserved interactions as well as conserved sequence. A second annotation method is unique to network alignment: If a protein of unknown function appears as part of an alignment together with a "landmark" protein of known function, we can use "landmark extension" to label the protein with a similar annotation. More highly connected and highly conserved alignments strengthen the hypothesis that the unknown protein shares function with the landmark protein.

Figure 6 shows an example of functional annotation through both pairwise and multiple network alignment. The pairwise alignment between *E. coli* and *C. crescentus* (Fig. 6A) shows a conserved DNA replication module. This includes the components of the primosome (dnaB, dnaA, gyrA, gyrB), the subunits of topoisomerase IV (parE, parC), and the β subunit of DNA polymerase III (dnaN). These protein families are all known to be involved in DNA replication.

Two aspects of this alignment are of interest. First, we see that the recF repair protein is linked to DNA replication in both organisms. Although this is not the primary annotation of the protein, a link to DNA replication was, in fact, found fairly recently (Kogoma 1997). Second, we observe the presence of the glucose-inhibited division proteins (gidA, gidB) and the protein trmE. Transcription of gidA affects DNA replication, both gidA and trmE are known to be involved in tRNA modification, and trmE has been implicated in cell cycle control (Gollop and March 1991). Taking these known interactions together with the alignment, we can hypothesize that both the gid proteins and trmE are likely to be involved in the cell-cycle-regulated control of DNA replication.

The multiple alignment diagram in Figure 6B extends the pairwise alignment to a multiple alignment of *E. coli*, *S. typhimurium*, *V. cholerae*, *C. crescentus*, *C. jejuni*, *H. pylori*, *M. tuberculosis*, *S. pneumoniae*, and *Synechocystis*. While some proteins from the pairwise alignment are absent, the core remains the same. The presence of the trmE protein in all nine species provides a compelling argument in favor of its role in DNA replication. This multiple alignment also offers another opportunity for landmark extension; the 60-kDa inner membrane protein yidC is present in all nine species and is highly connected to the other proteins in the alignment. Although known to be involved in protein secretion, the multiple alignment indicates that it is also likely to be linked to DNA replication.

Figure 7 shows an example of a 10-way multiple alignment relevant to bacterial cell division and cell envelope biogenesis. The alignment includes ftsZ, ftsW, and ftsI, well-known proteins

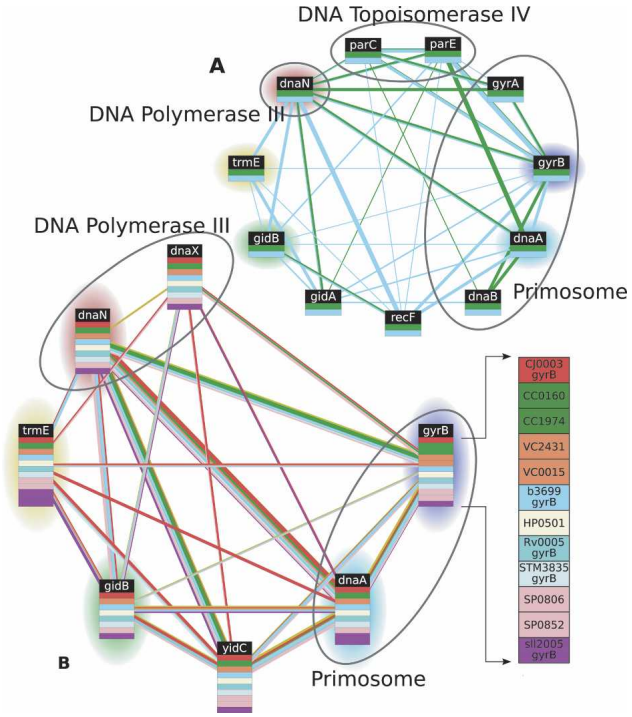

Figure 6. Two alignments of proteins involved in DNA replication. (*A*) A pairwise alignment between *E. coli* and *C. crescentus* includes several proteins involved in cell division as well as a conserved thiophene and furan oxidation protein. (*B*) A multiple alignment extends the pairwise alignment to include *S. typhimurium*, *V. cholerae*, *C. jejuni*, *H. pylori*, *M. tuberculosis*, *S. pneumoniae*, and *Synechocystis*. In this and subsequent figures, each colored box represents a protein and each vertical array of boxes represents an equivalence class; Græmlin hypothesizes that proteins in the same equivalence class performed the same function in the most recent common ancestor of the aligned species. To avoid clutter, individual proteins are not labeled, and, instead, each equivalence class is labeled with the consensus gene name of the proteins in it; as an example of the set of proteins aligned in an equivalence class, the detailed *inset* shows the specific proteins aligned to gyrB. Each protein is colored according to species, using the color code in Table 1; edges are also colored using the same scheme, and the width of each edge is proportional to its weight. In this figure, equivalence classes in the multiple alignment are highlighted the same color as the pairwise equivalence classes that they subsume.

involved in cell division, along with several other proteins from the *mur* and *mra* families known to be involved in peptidoglycan biogenesis. Many of these proteins are in contiguous operons in some species (Hara et al. 1997) but are scattered over the genome in species such as *C. jejuni* and *H. pylori*, rendering bioinformatics analysis difficult. This alignment, however, implicates them in cell division by association with the landmark proteins ftsZ, ftsW, and ftsI. In doing so, it uses information on the operon of one species (*E. coli*) to predict functional associations in the other species of the alignment.

Module identification

While support for functional annotation of proteins is currently the primary application of network alignment, the availability of numerous interaction networks may provide a resource for the study of functional modules. For example, Figure 8 shows that in *E. coli*, *S. typhimurium*, *V. cholerae*, *C. jejuni*, *H. pylori*, and *C. crescentus*, several proteins from the *exb*/*tol* family of biopolymer transporters are predicted to interact with a set of proteins involved in DNA recombination and integration. While the cooperation of these proteins is somewhat weak in any given species, the sum total of interactions in six distinct species suggests that DNA itself is the biopolymer transported through the *tol* channels before integrating into the chromosome. This alignment therefore may represent a part of a conserved module determining whether a cell is naturally competent for transformation (Dubnau 1999); this hypothesis is strengthened by studies showing that the insertional disruption of *exbB* in *Pseudomonas stutzeri* can reduce transformation efficiency to one-fifth of its previous level (Graupner and Wackernagel 2001). While in *P. stutzeri* the investigators used the fact that the *exb* genes were immediately downstream of two competence-related proteins, in species such as *C. jejuni* and *H. pylori* this chromosomal contiguity is not evident. Network alignment nevertheless identifies this module on the basis of conserved interactions.

Discussion

Interaction networks will soon constitute a vast store of data, as exemplified by the upcoming availability of hundreds of microbial interaction networks (Srinivasan et al. 2006). In light of this, network alignment is rapidly becoming an important analytical tool: Its goal is to map proteins of one interaction network to those of another and identify shared subnetworks that may constitute conserved biological modules. As with biosequence com-

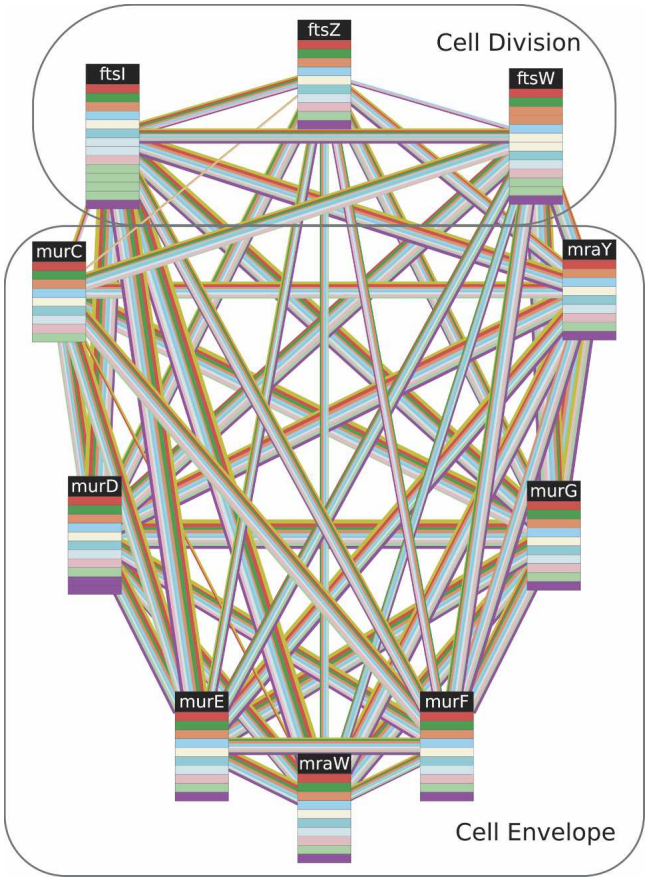

Figure 7. An alignment including proteins involved in cell division. This alignment implicates several proteins in bacterial cell division; it includes all species listed in Table 1.

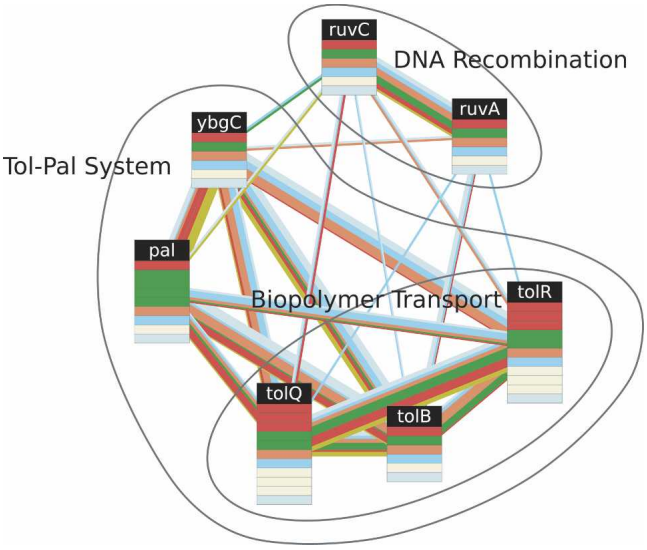

Figure 8. An alignment of a hypothetical functional module. In this alignment, proteins involved in biopolymer transport interact with proteins involved in DNA recombination. The sum total of these interactions in six species suggests that the proteins may be a part of a conserved functional module responsible for transformation.

parison, the principle that evolutionary conservation implies function can serve to increase the signal-to-noise ratio in analyses of interaction networks. With that help, biologists may be able to transfer functional annotations across species, extend annotations of known modules and landmark proteins to their strongly conserved neighboring proteins in a network, or identify novel modules by detecting unusual conserved subnetworks.

As our test results show, Græmlin is a promising method for network alignment. It scales efficiently to large inputs, particularly when searching databases, and is the first method capable of performing multiple alignments of an arbitrary number of networks. In addition, Græmlin uses a novel, flexible scoring scheme that can incorporate biologically trained parameters, and introduces the Module ESM framework, which offers the potential to search for subnetworks of arbitrary structure. In contrast, existing methods are limited to searching for multiprotein complexes, which are represented as fully connected graphs, or pathways, which are represented as ordered lists of proteins. As biological networks become less noisy and more complete, the Module ESM framework will allow fine-grained searches and analyses, and it will also offer the potential to refine models of known biological modules by quantifying the level of conservation of individual parts and interactions.

Our analyses of four conserved subnetworks accentuates several applications of network alignment, one of which is the analysis of proteins that lack functional annotation. Network alignment can do this by using conserved interactions and sequence to align an unknown protein to one with known function in another species. Alternatively, even if a protein has no homolog of known function, its occurrence as part of an alignment near well-known "landmark" proteins permits inferences about its function (Srinivasan et al. 2006).

As networks improve in quality and completeness, attention will focus on the functional annotation of modules in addition to proteins. Network alignment will play a key role by discovering groups of proteins that interact in more than one species, and it will thus offer additional evidence that such proteins work together to perform a common cellular function. As more networks become available, query-to-database network alignment will fulfill a similar role for modules as does BLAST for proteins (Altschul et al. 1997): By assembling a database of modules of known function, one may be able to annotate hypothetical modules that align to a module of known function in the database.

Although multiple network alignment is still in its infancy, it offers the potential to study modules in the context of functional evolution. Græmlin is a first step in the development of tools that will permit such studies, as it is capable of aligning many networks simultaneously and uses an evolutionarily based scoring scheme. Further algorithmic development will undoubtedly lead to data-motivated population genetic models for network evolution (McAdams et al. 2004; Koyuturk et al. 2005), where conserved interactions and conserved proteins will play the role of conserved residues. It is possible that even a SCOP-like hierarchy (Andreeva et al. 2004) for module families is on the horizon.

Although there is an extensive literature (Conte et al. 2004) on the topic of finding conserved graph topologies, the problems addressed by such algorithms are in general quite different from network alignment. For example, the evolutionary restriction on meaningful network alignments strongly constrains matches between graphs, as only homologous proteins from different species are aligned, whereas in the kind of graph matching treated by image processing algorithms (Conte et al. 2004), for example, nodes are tacitly assumed to be indistinguishable and edges represent indications of connectivity rather than beliefs about interaction. Another difference lies in the quality of the networks; probabilistic protein interaction networks are undirected graphs characterized by a low graph diameter (Barabasi and Oltvai 2004) and a high degree of topological uncertainty. As an extreme example of noisy graph structure, interaction networks based primarily on yeast two-hybrid data may not even be alignable, as several studies have questioned this assay's reliability (Bloom and Adami 2003; Drummond et al. 2005; Deeds et al. 2006). As networks increase in quality, however, ideas from general graph comparison techniques will be more relevant to network alignment.

With the impressive recent advances in sequencing, high-throughput techniques for gathering biological data, and computational methodologies for integrating such information into networks of protein interactions, comparisons of networks should become an increasingly important methodology for the molecular biologist. As our results show, Græmlin is a general and systematic methodology for comparing an arbitrary number of large networks. Many important challenges remain; for instance, the ability to reason about directed edges and align different types of interactions, such as physical contact and gene regulation, will allow more detailed analyses of biochemical pathways and regulatory cascades. On a more practical note, the ability to automatically identify interesting alignments for further study will be an important research topic unto itself.

Acknowledgments

J.F. was supported in part by a Stanford Graduate Fellowship; A.N. was supported in part by NLM training grant LM-07033 and NIH grant 5-T15-LM007033. J.F., A.N., and B.S.S. were funded by NSF grant EF-0312459, NIH grant UO1-HG003162, the NSF CAREER Award, and the Alfred P. Sloan Fellowship. B.S.S. was sup-

ported in part by a Department of Defense National Defense Science and Engineering Graduate Fellowship through the Army Research Office, and B.S.S. and H.H.M. were supported by NIH grant 1 R24 GM073011-01 and DOE Office of Science grant DE-FG02-01ER63219. We thank Andreas Sundquist for helpful comments on the manuscript.

References

Alexandersson, M., Cawley, S., and Pachter, L. 2003. SLAM: Cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Res.* **13:** 496–502.

Altschul, S.F., Carroll, R.J., and Lipman, D.J. 1989. Weights for data related by a tree. *J. Mol. Biol.* **207:** 647–653.

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

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T.J.P., Chothia, C., and Murzin, A.G. 2004. SCOP database in 2004: Refinements integrate structure and sequence family data. *Nucleic Acids Res.* **32:** D226–D229.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene Ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25:** 25–29.

Bafna, V. and Huson, D.H. 2000. The conserved exon method for gene finding. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8:** 3–12.

Barabasi, A.-L. and Oltvai, Z.N. 2004. Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.* **5:** 101–113.

Batzoglou, S. 2005. The many faces of sequence alignment. *Brief. Bioinform.* **6:** 6–22.

Batzoglou, S., Pachter, L., Mesirov, J.P., Berger, B., and Lander, E.S. 2000. Human and mouse gene structure: Comparative analysis and application to exon prediction. *Genome Res.* **10:** 950–958.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. 2004. Ultraconserved elements in the human genome. *Science* **304:** 1321–1325.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F.A., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14:** 708–715.

Bloom, J.D. and Adami, C. 2003. Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein–protein interactions data sets. *BMC Evol. Biol.* **3:** 21.

Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M., and Sherlock, G. 2004. GO::TermFinder—Open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20:** 3710–3715.

Bray, N. and Pachter, L. 2004. MAVID: Constrained ancestral alignment of multiple sequences. *Genome Res.* **14:** 693–699.

Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., Green, E.D., Sidow, A., and Batzoglou, S. 2003. LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13:** 721–731.

Chiaromonte, F., Yap, V.B., and Miller, W. 2002. Scoring pairwise genomic sequence alignments. *Pac. Symp. Biocomput.* 115–126.

Christie, K.R., Weng, S., Balakrishnan, R., Costanzo, M.C., Dolinski, K., Dwight, S.S., Engel, S.R., Feierbach, B., Fisk, D.G., Hirschman, J.E., et al. 2004. *Saccharomyces* Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res.* **32:** D311–D314.

Conte, D., Foggia, P., Sansone, C., and Vento, M. 2004. Thirty years of graph matching in pattern recognition. *IJPRAI* **18:** 265–298.

Cooper, G.M., Brudno, M., Stone, E.A., Dubchak, I., Batzoglou, S., and Sidow, A. 2004. Characterization of evolutionary rates and constraints in three mammalian genomes. *Genome Res.* **14:** 539–548.

Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S., and Sidow, A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15:** 901–913.

Dandekar, T., Schuster, S., Snel, B., Huynen, M., and Bork, P. 1999. Pathway alignment: Application to the comparative analysis of glycolytic enzymes. *Biochem. J.* **343:** 115–124.

Deeds, E.J., Ashenberg, O., and Shakhnovich, E.I. 2006. From The Cover: A simple physical model for scaling in protein–protein interaction networks. *Proc. Natl. Acad. Sci.* **103:** 311–316.

Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O., and Arnold, F.H. 2005. Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci.* **102:** 14338–14343.

Dubnau, D. 1999. DNA uptake in bacteria. *Annu. Rev. Microbiol.* **53:** 217–244.

Duda, R.O., Hart, P.E., and Stork, D.G. 2000. *Pattern classification*. Wiley-Interscience, New York.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. 1998. *Biological sequence analysis*. Cambridge University Press, UK.

Edgar, R.C. 2004. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5:** 113.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* **95:** 14863–14868.

Feng, D.F. and Doolittle, R.F. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* **25:** 351–360.

FlyBase Consortium. 2003. The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res.* **31:** 172–175.

Forst, C.V. and Schulten, K. 2001. Phylogenetic analysis of metabolic pathways. *J. Mol. Evol.* **52:** 471–489.

Fromont-Racine, M., Rain, J.C., and Legrain, P. 1997. Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat. Genet.* **16:** 277–282.

Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., et al. 2003. A protein interaction map of *Drosophila melanogaster*. *Science* **302:** 1727–1736.

Gollop, N. and March, P.E. 1991. A GTP-binding protein (Era) has an essential role in growth rate and cell cycle control in *Escherichia coli*. *J. Bacteriol.* **173:** 2265–2270.

Graupner, S. and Wackernagel, W. 2001. Identification and characterization of novel competence genes *comA* and *exbB* involved in natural genetic transformation of *Pseudomonas stutzeri*. *Res. Microbiol.* **152:** 451–460.

Hara, H., Yasuda, S., Horiuchi, K., and Park, J.T. 1997. A promoter for the first nine genes of the *Escherichia coli* mra cluster of cell division and cell envelope biosynthesis genes, including *ftsI* and *ftsW*. *J. Bacteriol.* **179:** 5802–5811.

Harris, T. W., Chen, N., Cunningham, F., Tello-Ruiz, M., Antoshechkin, I., Bastiani, C., Bieri, T., Blasiar, D., Bradnam, K., Chan, J., et al., 2004. WormBase: A multi-species resource for nematode biology and genomics. *Nucleic Acids Res.* **32:** D411–D417.

Hartwell, L.H., Hopfield, J.J., Leibler, S., and Murray, A.W. 1999. From molecular to modular cell biology. *Nature* **402:** 47–52.

Henikoff, S. and Henikoff, J.G. 1993. Performance evaluation of amino acid substitution matrices. *Proteins* **17:** 49–61.

Hillier, L.W., Miller, W., Birney, E., Warren, W., Hardison, R.C., Ponting, C.P., Bork, P., Burt, D.W., Groenen, M.A.M., Delany, M.E., et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432:** 695–716.

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

Kaneko, T., Tanaka, A., Sato, S., Kotani, H., Sazuka, T., Miyajima, N., Sugiura, M., and Tabata, S. 1995. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. I. Sequence features in the 1 Mb region from map positions 64% to 92% of the genome. *DNA Res.* **2:** 153–166.

Kelley, B.P., Sharan, R., Karp, R.M., Sittler, T., Root, D.E., Stockwell, B.R., and Ideker, T. 2003. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. Natl. Acad. Sci.* **100:** 11394–11399.

Kogoma, T. 1997. Is RecF a DNA replication protein? *Proc. Natl. Acad. Sci.* **94:** 3483–3484.

Korf, I., Flicek, P., Duan, D., and Brent, M.R. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17:** 140–148.

Koyuturk, M., Grama, A., and Szpankowski, W. 2005. Pairwise local alignment of protein interaction networks guided by models of evolution. *Lecture Notes in Bioinformatics* **3500:** 48–65.

Lee, I., Date, S.V., Adai, A.T., and Marcotte, E.M. 2004. A probabilistic functional network of yeast genes. *Science* **306:** 1555–1558.

Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.-O., Han, J.-D.J., Chesneau, A., Hao, T., et al. 2004. A map of the interactome network of the metazoan *C. elegans*. *Science* **303:** 540–543.

Lu, L.J., Xia, Y., Paccanaro, A., Yu, H., and Gerstein, M. 2005. Assessing the limits of genomic data integration for predicting protein networks. *Genome Res.* **15:** 945–953.

Ma, B., Tromp, J., and Li, M. 2002. PatternHunter: Faster and more

sensitive homology search. *Bioinformatics* **18:** 440–445.
Marchler-Bauer, A., Anderson, J.B., Cherukuri, P.F., DeWeese-Scott, C., Geer, L. Y., Gwadz, M., He, S., Hurwitz, D.I., Jackson, J.D., Ke, Z., et al. 2005. CDD: A Conserved Domain Database for protein classification. *Nucleic Acids Res*. **33:** D192–D196.
Matthews, L.R., Vaglio, P., Reboul, J., Ge, H., Davis, B.P., Garrels, J., Vincent, S., and Vidal, M. 2001. Identification of potential interaction networks using sequence-based searches for conserved protein–protein interactions or "interologs." *Genome Res*. **11:** 2120–2126.
McAdams, H.H., Srinivasan, B., and Arkin, A.P. 2004. The evolution of genetic regulatory systems in bacteria. *Nat. Rev. Genet*. **5:** 169–178.
Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. 2002. Network motifs: Simple building blocks of complex networks. *Science* **298:** 824–827.
Ogata, H., Fujibuchi, W., Goto, S., and Kanehisa, M. 2000. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Res*. **28:** 4021–4028.
Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O. 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci*. **96:** 4285–4288.
Pinter, R.Y., Rokhlenko, O., Yeger-Lotem, E., and Ziv-Ukelson, M. 2005. Alignment of metabolic pathways. *Bioinformatics* **21:** 3401–3408.
Sharan, R., Ideker, T., Kelley, B., Shamir, R., and Karp, R.M. 2005a. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *J. Comput. Biol*. **12:** 835–846.
Sharan, R., Suthram, S., Kelley, R.M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R.M., and Ideker, T. 2005b. Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci*. **102:** 1974–1979.
Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. **15:** 1034–1050.
Srinivasan, B.S., Novak, A., Flannick, J., Batzoglou, S., and McAdams, H.H. 2006. Integrated protein interaction networks for 11 microbes. In *Proceedings of the 10th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2006)* (in press).
Stuart, J.M., Segal, E., Koller, D., and Kim, S.K. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302:** 249–255.
Sun, Y. and Buhler, J. 2005. Designing multiple simultaneous seeds for DNA similarity search. *J. Comput. Biol*. **12:** 847–861.
Tatusov, R.L., Koonin, E.V., and Lipman, D.J. 1997. A genomic perspective on protein families. *Science* **278:** 631–637.
Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., et al. 2000. A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403:** 623–627.
Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.-M., and Eisenberg, D. 2002. DIP, the Database of Interacting Proteins: A research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*. **30:** 303–305.

Received February 17, 2006; accepted in revised form May 18, 2006.