Chapter 3

Global Alignment of Protein–Protein Interaction Networks

Misael Mongiovì and Roded Sharan

Abstract

Sequence-based comparisons have been the workhorse of bioinformatics for the past four decades, furthering our understanding of gene function and evolution. Over the last decade, a plethora of technologies have matured for measuring Protein–protein interactions (PPIs) at large scale, yielding comprehensive PPI networks for over ten species. In this chapter, we review methods for harnessing PPI networks to improve the detection of orthologous proteins across species. In particular, we focus on pairwise global network alignment methods that aim to find a mapping between the networks of two species that maximizes the sequence and interaction similarities between matched nodes. We further suggest a novel evolutionary-based global alignment algorithm. We then compare the different methods on a yeast-fly-worm benchmark, discuss their performance differences, and conclude with open directions for future research.

Key words: Network alignment, Protein–protein interaction, Functional orthology, Network evolution

1. Introduction

Over the last decade, high-throughput techniques such as yeast two-hybrid assays (1) and co-immunoprecipitation experiments (2), have allowed the construction of large-scale networks of Protein–protein interactions (PPIs) for multiple species. Comparative analyses of these networks have greatly enhanced our understanding of protein function and evolution.

Analogously to the sequence comparison domain, two main concepts have been introduced in the network comparison context: *local* network alignment and *global* network alignment. The first considers local regions of the network, aiming to identify small subnetworks that are conserved across two or more species (where conservation is measured in terms of both sequence and interaction patterns). Local alignment algorithms have been utilized to detect

Hiroshi Mamitsuka et al. (eds.), *Data Mining for Systems Biology: Methods and Protocols*, Methods in Molecular Biology, vol. 939, DOI 10.1007/978-1-62703-107-3_3,

protein pathways (3) and complexes that are conserved across multiple species (4–6), to predict protein function, and to infer novel PPIs (4).

In global network alignment (GNA), the goal is to associate proteins from two or more species in a global manner so as to maximize the rate of sequence and interaction conservation across the aligned networks. In its simplest form, the problem calls for identifying a 1-1 mapping between the proteins of two species so as to optimize some conservation criterion. Extensions of the problem consider multiple networks and many-to-many (rather than 1-1) mappings. Such analyses assist in identifying (functional) orthologous proteins and orthology families (7) with applications to predicting protein function and interaction. They aim to improve upon sequence-only methods that partition proteins into orthologous groups based on sequence-similarity computations (8–10).

GNA methods can be classified into two main categories. The first category contains matching methods that explicitly search for a one-to-one mapping that maximizes a suitable scoring function. The scoring function favors mappings that conserve sequence and interaction. Methods in this category include the integer linear programming (ILP) method of (11) and a greedy gradient ascent method of (12). The second category includes ranking methods that consider all possible pairs of interspecies proteins that are sufficiently sequence-similar, and rank them according to their sequence and topological similarity. These ranks are then used to derive a 1-1 mapping. Methods in this category include a Markov random field (MRF) approach (13), the IsoRank method that is based on Google's Page Rank (7), and a diffusion-based method—hybrid RankProp (14). In addition, there are several very recent ranking approaches that do not use sequence-similarity information at all (15, 16).

Here, we aim to propose a third, evolutionary perspective on global alignment by designing a GNA algorithm that is based on a probabilistic model of network evolution. The evolution of a network is described in terms of four basic events: gene duplication, gene loss, edge attachment, and edge detachment. This model allows the computation of the probability of observing extant networks given the ancestral network they originated from; by maximizing this probability, one obtains the most likely ancestor–descendant relations, which naturally translate into a network alignment.

This chapter is organized as follows: Subheading 3 reviews GNA methods that are based on graph matching. Subheading 4 presents the ranking-based methods. Subheading 5 describes in detail the probabilistic model of evolution and the proposed alignment method. The different approaches are compared in Subheading 6. Finally, Subheading 7 gives a brief summary and discusses future research directions.

2. Preliminaries and Problem Definition

We focus the presentation on methods for pairwise global alignment, where the input consists of two networks and possibly sequence-similarity information between their nodes, and the output is a correspondence, commonly one-to-one, between the nodes of the two networks.

A protein network $G=(V, E)$ has a set V of nodes, corresponding to proteins, and a set E of edges, corresponding to PPIs. For a node $i \in V$, we denote its set of (direct) neighbors by $N(i)$. Let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ be the two networks to be aligned. Let $R \subseteq V_1 \times V_2$ be a compatibility relation between proteins of the two networks, representing pairs of proteins that are sufficiently sequence-similar. A many-to-many correspondence that is *consistent* with R is any subset $R^* \subseteq R$. Under such a correspondence, we say that an edge (u, v) in one of the networks is *conserved* if there exists an edge (u', v') in the other network such that $(u, u'), (v, v') \in R^*$ or $(u', u), (v', v) \in R^*$. We let $T(G_1, G_2) = \{(u, u', v, v'): (u, v), (u', v') \in R, (u, u') \in E_1, (v, v') \in E_2\}$ denote the set of all quadruples of nodes that induce a conserved interaction.

In its simplest formulation, the alignment problem is defined as the problem of finding an injective function (one-to-one mapping) φ: $V_1 \rightarrow V_2$ such that (i) it is consistent with R and (ii) it maximizes the number of conserved interactions. More elaborate formulations of the problem can relax the 1-1 mapping to a many-to-many mapping and possibly define an alignment score to be optimized that combines the amount of interaction conservation and the sequence similarity of the matched nodes. The definition of a conserved interaction can also be made more elaborate by taking into account the reliability of the pertaining interactions and by allowing "gapped" interactions, i.e., a directed interaction in one network is matched to two nodes that are of distance 2 in the other network. We defer the discussion of these extensions and the specific scoring functions used to the next sections, where the different GNA approaches are described.

The problem of finding the optimal one-to-one alignment between two networks, as defined above, can be shown to be NP-hard by reduction from maximum common subgraph (11). Consequently, an efficient algorithm cannot be designed for the general case. However, under certain relaxations the problem can be solved optimally on current data sets in acceptable time.

3. Graph Matching Methods

In this section, we describe GNA methods that look for an explicit 1-1 correspondence between the two compared networks. The first method, by Klau, is based on reformulating the alignment problem

as an ILP (11). The variables of the program represent the 1-1 mapping sought. Specifically, for each pair $(u, v) \in R$, the author defines a binary variable x_{uv} denoting whether u and v are matched ($x_{u,v} = 1$) in the alignment or not ($x_{u,v} = 0$). The ILP formulation is as follows:

$$\max \sum_{(u,u',v,v') \in T(G_1,G_2)} x_{u,v} \cdot x_{u',v'} + \sum_{(u,v) \in R} \sigma(u, v) \cdot x_{u,v}$$

s.t.

$$\sum_{u \in V_1} x_{u,v} \leq 1 \quad \forall v \in V_2$$

$$\sum_{v \in V_2} x_{u,v} \leq 1 \quad \forall u \in V_1,$$

where $\sigma(u, v)$ denotes the sequence similarity of u and v. The objective function can be linearized in an obvious way by introducing binary variables $t_{u,u',v,v'} = x_{u,v} \cdot x_{u',v'}$ (for $(u, u', v, v') \in T(G_1, G_2)$) with appropriate constraints.

While the author uses optimization techniques, such as Lagrangian decomposition and Lagrangian relaxation, to solve this problem, an optimum solution for restricted instances can be found in reasonable time as we report in Subheading 6. We note that if $V_1 \cap V_2$ is first partitioned into sufficiently small orthology clusters (using, e.g., the Inparanoid algorithm (8)) and if the graph of potential conserved interactions across clusters has no loops, then the optimum alignment can be found in polynomial time via a dynamic programming algorithm (12).

In the general case, the computation of optimal solutions is too costly, hence the use of heuristics is necessary. Vert et al. (12) suggested a gradient ascent approach. It starts from a feasible solution and computes a sequence of moves in the direction of the objective's gradient until converging to a local maximum. Denoting the adjacency matrices of the two graphs by A_1 and A_2, respectively, and assuming that $|V_1| = |V_2| = n$ (otherwise, add dummy vertices), the goal of the optimization is to find a permutation matrix P that maximizes a weighted sum of the number $J(P)$ of conserved interactions and a sequence similarity term $S(P)$. In matrix notation, $J(P) = \frac{1}{2} tr(A_1^T P A_2 P^T)$ and its gradient is $A_1^T P A_2$; $SP = tr(PC)$ where C, the matrix of sequence-similarity scores, is its gradient.

The initial solution P_0 is given by sequence similarity alone, using a maximum matching algorithm. At each step, the algorithm employs a maximum matching computation to update the current permutation in the direction of the gradient:

$$P_{n+1} = \arg\max_P tr([\lambda A_1^T P_n A_2 + (1 - \lambda) C] P),$$

where $0 \leq \lambda \leq 1$ is a weighting constant.

4. Methods Based on Ranking

A second class of methods is based on assigning a score to each pair of compatible nodes and only at a second step choosing a global pairing of the nodes. The latter pairing is effectively *disambiguating* the compatibility relations, pinpointing the "best" 1-1 mapping. The disambiguation can be achieved by computing a maximum weighted bipartite matching or via simple greedy strategies. The difference between the various methods lies mainly in the first, scoring phase.

The first method for GNA has been proposed by Bandyopadhyay et al. (13) and uses a ranking that is based on a MRF model. It starts by building an alignment graph, where the nodes represent candidate pairs of (sequence-similar) proteins and the edges represent potentially conserved interactions. Each node in the alignment graph is associated with a binary state z indicating if that node represents a true orthology relation or not. The state values are modeled using a MRF. The MRF model assumes that for each node of the alignment graph $j = (u, v)$, the probability that j represents a true pair of orthologs ($z_j = 1$) depends only on the states of its neighbors ($N(j)$), and the dependence is through a logistic function:

$$P(z_j | z_{N(j)}) = \frac{1}{1 + e^{-\alpha - \beta \cdot c(j)}},$$

where α and β are parameters and $c(j)$ is the conservation index of j, defined as twice the number of conserved interactions between j and neighbors of j whose states are pre-assigned with value 1 (true orthologs), divided by the total number of interactions involving u and v across the two species. The inference of the states of the nodes is conducted using Gibbs sampling (17), yielding orthology probabilities for every node. These estimated probabilities are used to disambiguate the pairing.

Singh et al. (7) proposed an alignment method (IsoRank) that is based on Google's PageRank algorithm. As for MRF, the method first computes a score for each candidate pair of orthologs and then uses the scores for disambiguating the pairing. The score $R_{(i,j)}$ of the pair $(i, j) \in V_1 \times V_2$ is a weighted average of the scores of its neighboring pairs (assuming that all node pairings are allowed):

$$R_{(i,j)} = \sum_{u \in N(i)} \sum_{v \in N(j)} \frac{R_{(u,v)}}{|N(u)||N(v)|}.$$

The authors translate the problem of finding R into an eigenvector problem by expressing it in matrix form as $R = AR$ where A is defined as:

$$A_{(i,j)(u,v)} = \begin{cases} \frac{1}{|N(u)||N(v)|} & \text{if } (i,u) \in E_1, (j,v) \in E_2 \\ 0 & \text{otherwise.} \end{cases}$$

Under this formulation, the problem reduces to finding the dominant eigenvector of A, which is efficiently solved using the power method. To account for sequence similarity, the objective is modified as $R = [\alpha A + (1-\alpha)B1^T]R$ where B is the vector of normalized bit scores and 1^T is an all-1 row vector.

Yosef et al. (14) devised the hybrid RankProp algorithm. It considers one "query node" of the first network at a time and ranks the nodes of the second network with respect to it by using a diffusion procedure. To this end, they constructed a composite network with two types of edges: PPI and sequence similarity. The query node is assigned a score of 1.0 that is continually pumped to the other nodes through the network's edges. The scores that the nodes assume after the diffusion process converges induce a ranked list of candidates for matching the query node. In detail, at step $t + 1$, the score of a node i with respect to a query q is given by:

$$S_i(t+1) = W_{qi} + \alpha \sum_{j \in N(i) \setminus \{q\}} W_{ji} S_j(t),$$

where α is a parameter controlling the diffusion rate and W is a weight matrix that represents the composite network—it is the normalized confidence of an interaction for PPI edges and a normalized sequence similarity for sequence-similarity edges. Finally, to make the score symmetric, proteins from both networks are queried and each pair is assigned the average score of its two associated queries.

5. Network Evolution-Based Alignment

In this section, we present a new alignment method, called PME, that is based on a probabilistic model of evolution. PME aims to reconstruct the most probable ancestral network that gave rise to the observed extant networks. Such a network induces a many-to-many alignment in the descending networks by associating groups of proteins in the two input networks with the corresponding ancestral proteins. The method is based on a probabilistic model of the evolutionary dynamics of a network, that supports four kinds of evolutionary events: link attachment, link detachment, gene duplication and gene loss (18).

An alignment between two networks G_1 and G_2 is defined by an ancestral network $G_0 = (V_0, E_0)$ and two functions $f_1 : V_1 \rightarrow V_0$ and

$f_2 : V_2 \rightarrow V_0$ which map the nodes of G_1 and G_2 into the nodes of G_0 (ancestral proteins). The score of an alignment $A = (G_0, f_1, f_2)$ is the product of the prior probability for A and the likelihood of observing G_1 and G_2 given A. We describe the probability computations in detail below.

The probability $P(A)$ is the product of two terms that consider the prior probability of observing G_0 and the probability of the pattern of gene duplications and losses implied by f_1 and f_2. For the former, we adopt a simple Erdős–Rényi model where edges occur independently with some constant probability P_E. For the latter, we focus on gene duplications (as in (18)), assuming that gene duplication events occur independently with some fixed probability P_d. For computational efficiency, we disallow gene losses, although those could be easily incorporated to the model in a similar manner. Formally, the two terms are as follows:

- A priori ancestral network probability:

$$\prod_{(u,v)\notin E_0} (1 - P_E) \cdot \prod_{(u,v)\in E_0} P_E.$$

- Gene duplication ($i \in \{1, 2\}$):

$$\prod_{\substack{v\in V_0 \\ f_i^{-1}(v)\neq\phi}} P_d^{|f_i^{-1}(v)|-1} \cdot \prod_{\substack{v\in V_0 \\ |f_i^{-1}(v)|\leq 1}} (1 - P_d).$$

The probability $P(G_i|A)$ of observing the network G_i, $i \in \{1, 2\}$ is given by the product of two factors that consider edge attachment and edge detachment events, assuming these events occur independently with probabilities P_A and P_D, respectively.

- Edge attachment:

$$\prod_{(u,v)\notin E_0} \left(\prod_{\substack{(u',v')\notin E_i \\ f_i(u')=u, f_i(v')=v}} (1 - P_A) \cdot \prod_{\substack{(u',v')\in E_i \\ f_i(u')=u, f_i(v')=v}} P_A \right).$$

- Edge detachment:

$$\prod_{(u,v)\in E_0} \left(\prod_{\substack{(u',v')\in E_i \\ f_i(u')=u, f_i(v')=v}} (1 - P_D) \cdot \prod_{\substack{(u',v')\notin E_i \\ f_i(u')=u, f_i(v')=v}} P_D \right).$$

Our goal is to find an alignment that maximizes $P(G_1, G_2, A) = P(A) \cdot P(G_1|A) \cdot P(G_2|A)$. In the following, we provide an ILP

formulation of the problem. Consider a set of n hypothetical nodes of the ancestral network, where $n = |V_1| + |V_2|$ is the maximal number of nodes in the ancestral network. With each node, we associate a binary variable z_i which is 1 if and only if node i has some descendant node in the extant networks. With each vertex pair (i, j), we associate a binary variable t_{ij} which is 1 if and only if nodes i and j interact with each other in the ancestral network. To model the mappings f_1 and f_2, we define binary variables x_{iu} and y_{iv}, where $x_{iu} = 1$ ($y_{iv} = 1$) if and only if $f_1(u) = i$ ($f_2(v) = i$). Finally, in order to consider gene duplications, we add binary variables d_i^j, $j \in \{1, 2\}$ such that $d_i^j = 0$ if and only if i has more than one descendant in G_j.

5.1. The ILP Formulation

The constraints of the ILP are defined as follows:

$$t_{ij} \leq z_i, z_j, \quad 1 \leq i < j \leq n$$

to allow edges only between "true" vertices of the ancestral network.

$$\sum_{i=1}^{n} x_{iu} = 1, \quad u \in V_1,$$

$$\sum_{i=1}^{n} y_{iv} = 1, \quad \in V_2$$

to model the fact that each protein descends from a single ancestor.

$$\sum_{u \in V_1} x_{iu} \geq z_i, \quad 1 \leq i \leq n,$$

$$\sum_{v \in V_2} y_{iv} \geq z_i, \quad 1 \leq i \leq n,$$

$$x_{iu} \leq z_i, \quad 1 \leq i \leq n, \quad u \in V_1,$$

$$y_{iu} \leq z_i, \quad 1 \leq i, \leq n, \quad u \in V_2$$

to model the fact that each true node of the ancestral network ($z_i = 1$) must have at least one descendant in each network and each dummy node of the ancestral network ($z_i = 0$) cannot have any descendants.

$$d_i^1 \leq 1 + z_i - x_{iu} - x_{iv}, \quad 1 \leq i \leq n, \quad u, v \in V_1,$$

$$d_i^1 \geq 1 + z_i - \sum_{u \in V_1} x_{iu}, \quad 1 \leq i \leq n,$$

$$d_i^2 \leq 1 + z_i - y_{iu} - y_{iv}, \quad 1 \leq i \leq n, \quad u, v \in V_2,$$

$$d_i^2 \geq 1 + z_i - \sum_{u \in V_2} y_{iu}, \quad 1 \leq i \leq n$$

to impose that nodes that have only one descendant have not undergone a duplication event. Finally, we add the integer constraints:

$$x_{iu}, y_{iv}, z_i, t_{ij}, d_i^1, d_i^2 \in \{0, 1\} \; 1 \leq i, j \leq n, u \in V_1, v \in V_2.$$

The objective is to maximize $P(G_1, G_2, A)$ or, equivalently, to maximize $\log P(G_1, G_2, A)$. The latter is a sum of four terms:

- A priori ancestral network probability:

$$\varphi_E = \sum_{i<j} \big(\log(P_E) \cdot t_{ij} + \log(1 - P_E) \cdot (1 - t_{ij})\big).$$

- Gene duplication (for simplicity, we specify only the sub-term involving G_1):

$$\varphi_d = \sum_{i=1}^{n} \left(\sum_{u \in V_1} x_{iu} - z_i \right) \cdot \log(P_d) + \sum_{i=1}^{n} \log(1 - P_d) \cdot d_i^1.$$

- Edge attachment (for simplicity, we specify only the sub-term involving G_1):

$$\varphi_A = \sum_{i<j} (1 - t_{ij}) \cdot \left(\sum_{(u,v) \notin E_1} x_{iu} \cdot x_{jv} \cdot \log(1 - P_A) + \sum_{(u,v) \in E_1} x_{iu} \cdot x_{jv} \cdot \log(P_A) \right).$$

- Edge detachment (for simplicity, we specify only the sub-term involving G_1):

$$\varphi_D = \sum_{i<j} t_{ij} \cdot \left(\sum_{(u,v) \in E_1} x_{iu} \cdot x_{jv} \cdot \log(1 - P_D) + \sum_{(u,v) \notin E_1} x_{iu} \cdot x_{jv} \cdot \log(P_D) \right).$$

In order to make the problem linear, we introduce the following additional binary variables with appropriate constraints: $p_{ijuv} = t_{ij} \cdot x_{iu} \cdot x_{jv}$ and $q_{ijuv} = (1 - t_{ij}) \cdot x_{iu} \cdot x_{jv}$.

5.2. Refinements and Variable Reduction

In some cases, there are not enough interactions to support a match. To avoid an arbitrary choice among identically scored solutions, we choose the solution that agrees best with the sequence-similarity information. To this end, we add a small penalty to each ancestral-descendant connection whose value is $10^{-8} \cdot \log(S + 1)$, where S is the bit score of the two proteins.

Although PME naturally produces a many-to-many correspondence between orthologous proteins, we focus here on its reduction to a one-to-one mapping to facilitate its comparison to other methods. To this end, we rank all pairs of inter-species proteins that are predicted to descend from the same common ancestor. For any potentially

matched pair (u, v), with $f(u) = f(v) = i$, the score of (u, v) is given by the score of the global alignment after removing all the nodes that descend from i except for u and v (i.e., forcing the alignment to match u and v). These scores are then fed to a maximum bipartite matching computation to construct a 1-1 alignment.

The sequence-similarity information allows us to greatly reduce the number of variables considered. We start with a set $V = V_1 \cap V_2$ of hypothetical ancestral nodes. We build two relations $R_1 \subseteq V \times V_1$ and $R_2 \subseteq V \times V_2$ as follows: For each $i \in V$, we add to R_1 all pairs (i, u) with $u \in V_1$ such as u is sequence-similar to i and $u \leq i$. Analogously, we add to R_2 all pairs (i, v) with $v \in V_2$ such that u is sequence-similar to i and $u \leq i$. The search is then restricted to alignments whose ancestor–descendant pairs are in $R_1 \cap R_2$.

The relations R_1 and R_2 also allow us to reduce the number of possible edges of the ancestral network. Consider a pair of nodes (u, v) of the ancestral network such that all possible pairs of descendants of these nodes span non-edges. Clearly, in the optimal solution, (u, v) will be a non-edge. Since the networks are usually very sparse, this simple rule greatly reduces the number of variables required to model the topology of the ancestral network and, consequently, greatly saves in variables introduced by the linearization. Although non-edges contribute to the objective function, we can modify the latter so that the contribution of non-edges is zero (by adding $-\log(1 - P_E)$ to all ancestral vertex pairs). In a similar manner, we can reduce the number of ancestor–descendant pairs that are considered in the computation of edge attachment events.

6. Experimental Results

To compare the different GNA methods, we used the benchmark in (13), which focuses on the pairwise global alignment of the PPI networks of yeast and fly, starting from an initial clustering of the proteins into orthology families formed by the Inparanoid algorithm (8). In addition, we compared, under the same setting, the alignments of each of these networks to a PPI network of worm. The worm network was constructed by collecting data from recently published papers and public databases (19–21) and spanned 2,967 proteins and 4,852 interactions. The yeast network contained 4,393 proteins and 14,318 interactions; the fly network contained 7,042 proteins and 20,719 interactions. We considered 2,244 Inparanoid groups between yeast and fly, 1,833 groups between yeast and worm, and 4,228 groups between worm and fly.

We included in the comparison the following methods: ILP (11), MRF (13), IsoRank (7) and PME (Subheading 5). We did not consider gradient ascent (12) and hybrid RankProp (14)

in our tests. Gradient ascent tries to approximate the same objective as the ILP method, hence the latter should be superior to it. Hybrid RankProp was shown by its authors to be equivalent in performance to the original RankProp method, which is based on sequence only.

We implemented ILP, IsoRank, and PME in Matlab and used ILOG CPLEX as an ILP solver. For MRF, we report on the results published in the original paper (13). The parameter that balance topology versus sequence similarity was set as $\epsilon = 0.01$ for both IsoRank and ILP in order to give higher weight to topology. For PME, we used the following settings: The probability of attachment and detachment was set so as to obtain the same global rate of attachments and detachments estimated from the unambiguous clusters of Inparanoid ($P_A = 0.0026$; $P_D = 0.9617$). The probability of an edge in the ancestral network was estimated from the density of the two networks ($P_E = 3.32\epsilon^{-4}$). The probability of duplication was set to $P_d = 0.03$ with the results being robust to a wide range of values for this parameter (in the range 10^{-4}–0.5). All the experiments were executed on a DELL server with eight processors Quad-Core AMD Opteron and 16 GB RAM, OS Ubuntu 9.04.

To evaluate the functional coherency of the aligned proteins, we considered two measures: (1) the number of pairs that are classified as orthologs by HomoloGene (22), considered as a gold standard; and (2) a score based on the gene ontology (GO) (23), focusing on the biological process and molecular function branches. To evaluate the significance of the number of HomoloGene pairs that were matched, we computed a hypergeometric p-value, which measures the probability that a random set of matches (of the same size as our alignment and constrained to the Inparanoid clusters) would yield the observed overlap or higher. The GO score is computed as the average GO similarity of all matched pairs. We employed the Resnik similarity among terms and considered as a similarity between proteins the value of the best matching between their terms (24). We restricted our analysis to the set of ambiguous clusters, i.e., clusters that contain more than one protein for at least one of the species.

The results for yeast-fly, yeast-worm and fly-worm are reported in Table 1. Evidently, all methods perform similarly. ILP and IsoRank always attain the maximum number of conserved interactions. This is expected for ILP and suggests that IsoRank is a good heuristic for maximizing the number of conserved interactions. ILP also achieves the maximum number of HomoloGene pairs, except in the yeast-worm alignment, where it is outperformed by IsoRank. With respect to the GO measures, ILP attains the highest scores in most cases, with PME performing better on the molecular function measure.

Table 1
A comparison of GNA methods on a yeast-fly-worm benchmark

Dataset	Method	Total pairs	Conserved interactions	HomoloGene			GO Sim	
				Pairs	%	P-value	(MF)	(BP)
Yeast-fly	ILP	545	91	134	0.246	4.33e − 09	3.32	1.87
	MRF	535	87	133	0.248	2.17e − 09	3.26	1.85
	IsoRank	545	91	133	0.244	1.01e − 08	3.28	1.88
	PME	545	86	132	0.242	2.33e − 08	3.25	1.83
Yeast-worm	ILP	194	48	72	0.371	0.059	2.95	2.23
	IsoRank	194	48	74	0.381	0.021	2.97	2.22
	PME	194	47	72	0.371	0.059	2.98	2.22
Fly-worm	ILP	209	38	93	0.445	0.004	2.32	1.62
	IsoRank	209	38	87	0.416	0.084	2.32	1.50
	PME	209	36	92	0.440	0.007	2.34	1.61

7. Conclusions

In this chapter, we present the GNA problem and discuss extant methods for solving it. A guiding principle in most of these methods is the maximization of conserved interactions across the two aligned networks. We further present a novel strategy to the problem that is based on a probabilistic model of protein network evolution. We test the methods on a yeast-fly-worm benchmark and find that all methods perform similarly on current networks when starting from a defined set of orthology groups.

We believe that future research in this domain should cover both the development of better alignment methods and the benchmarking of such methods. While current methods do reasonably well with respect to maximizing the number of conserved interactions, evolutionary considerations are still scarcely used and could potentially guide the alignment in a more refined way, particularly when comparing species that are less distant apart. Additional developments could include going beyond 1-1 alignments and pairwise comparisons (25). An orthogonal axis is the development of gold standard alignments. Current benchmarks such as the Homologene collection are mostly sequence-driven and, thus, potentially lead to biased assessment of methods. In summary, we expect GNA methods to have greater impact as protein networks and orthology information continue to accumulate.

Acknowledgements

M.M. was partially supported by the Army Research Laboratory, under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the US Government. The US Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation here on. R.S. was supported by a research grant from the Israel Science Foundation (grant no. 385/06).

References

1. Fields S, Song O (1989) A novel genetic system to detect Protein–protein interactions. Nature 340(6230):245–246
2. Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. Nature 422:198–207
3. Kelley BP, Yuan B, Lewitter F, Sharan R, Stockwell BR, Ideker T (2004) PathBLAST: a tool for alignment of protein interaction networks. Nucl Acids Res 32(Suppl 2): W83–W88
4. Sharan R, Suthram S, Kelley R, Kuhn T, McCuine S, Uetz P, Sittler T, Karp R, Ideker T (2005) Conserved patterns of protein interaction in multiple species. Proc Natl Acad Sci USA 102(6):1974–1979
5. Kalaev M, Bafna V, Sharan R (2009) Fast and accurate alignment of multiple protein networks. J Comput Biol 16(8):989–999
6. Koyuturk M, Kim Y, Topkara U, Subramaniam S, Szpankowski W, Grama A (2006) Pairwise alignment of protein interaction networks. J Comput Biol 13(2):182–199
7. Singh R, Xu J, Berger B (2008) Global alignment of multiple protein interaction networks with application to functional orthology detection. Proc Natl Acad Sci USA 105 (35):12763–12768
8. Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and inparalogs from pairwise species comparisons. J Mol Biol 314(5):1041–1052
9. Tatusov R et al (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 1(4):41
10. Datta RS, Meacham C, Samad B, Neyer C, Sjolander K (2009) Berkeley PHOG: phyloFacts orthology group prediction web server. Nucl Acids Res 37(Suppl 2):84–89
11. Klau G (2009) A new graph-based method for pairwise global network alignment. BMC Bioinformatics 10(Suppl 1):S59
12. Zaslavskiy M, Bach F, Vert JP (2009) Global alignment of Protein–protein interaction networks by graph matching methods. Bioinformatics 25(12):i259–1267
13. Bandyopadhyay S, Sharan R, Ideker T (2006) Systematic identification of functional orthologs based on protein network comparison. Genome Res 16(3):428–435
14. Yosef N, Sharan R, Noble WS (2008) Improved network-based identification of protein orthologs. Bioinformatics 24(16):i200–i206
15. Milenkovic T, Wong W, Hayes W, Przulj N (2010) Optimal network alignment with graphlet degree vectors. Canc Inform 9:121–137
16. Kuchaiev O, Milenkovic T, Memisevic V, Hayes W, Przulj N (2010) Topological network alignment uncovers biological function and phylogeny. J R Soc Interface 7(50):1341–1354
17. Smith AFM, Roberts GO (1993) Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. J Roy Stat Soc B Stat Meth 55(1):3–23
18. Berg J, Lassig M, Wagner A (2004) Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. BMC Evol Biol 4(1):51
19. Li S et al (2004) A map of the interactome network of the Metazoan C. elegans. Science 303(5657):540–543
20. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D (2002) DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. Nucl Acids Res 30(1):303–305
21. Chen N et al (2005) WormBase: a comprehensive data resource for Caenorhabditis biology

and genomics. Nucl Acids Res 33(Suppl 1): D383–389

22. Wheeler DL et al (2005) Database resources of the national center for biotechnology information. Nucl Acids Res 33(Suppl 1):D39–D45
23. Ashburner M (2000) Gene ontology: tool for the unification of biology. Nat Genet 25:25–29
24. Schlicker A, Albrecht M (2007) FunSimMat: a comprehensive functional similarity database. Nucl Acids Res 36(Suppl 1):D434–439
25. Liao CS, Lu K, Baym M, Singh R, Berger, B (2009) IsoRankN: spectral methods for global alignment of multiple protein networks. Bioinformatics 25(12):i253–258