A Novel Algorithm for Local Alignment of Protein Interaction Networks: MODULA

Extended Abstract

Pietro H Guzzi[1], Pierangelo Veltri[1], Swarup Roy[2], and Jugal K Kalita[2]

[1] Dept Surgical Medical Sciences Unicz, Italy,
hguzzi,veltri@unicz.it
[2] Department of Information Technology, North-Eastern Hill University,India
swarup@nehu.ac.in
[3] Department of Computer Science
University of Colorado, Colorado Springs, USA
jkalita@uccs.edu

Abstract. Biological networks are usually used to model interactions among biological macromolecules in a cells. For instance protein-protein interaction networks (PIN) are used to model and analyse the set of interactions among proteins. The comparison of networks may result in the identification of conserved patterns of interactions corresponding to biological relevant entities such as protein complexes and pathways. Several algorithms, known as network alignment algorithms, have been proposed to unravel relations between different species at the interactome level. Algorithms may be categorized in two main classes: merge and mine and mine and merge. Algorithms belonging to the first class initially merge input network into a single integrated and then mine such networks. Conversely algorithms belonging to the second class initially analyze separately two input networks then integrate such results. In this paper we present MODULA (Network **Modul**e based PPI **A**ligner), a novel approach for local network alignment that belong to the second class. The algorithm at first identifies compact modules from input networks. Modules of both networks are then matched using functional knowledge. Then it uses high scoring pairs of modules as seeds to build a bigger alignment. In order to asses MODULA we compared it to the state of the art local alignment algorithms over a rather extensive and updated dataset. [4]

1 Introduction

Complex biological systems are often represented as networks and studied computationally. In PPI networks [1, 2], also known as protein interaction networks (PINs), the proteins are represented by nodes and interactions between them are represented by edges. Studies suggest that molecular networks are conserved through evolution [3, 4], and that highly connected proteins are more likely to

[4] This work has been presented in IEEE BIBM 2015.

be essential for survival than proteins with lower connectivity. As a result, the interactions between protein pairs as well as the overall composition of the network are important for the overall functioning of an organism. Understanding conserved substructures through comparative analysis of these networks can provide basic insights into a variety of biochemical processes. The ultimate goal of network alignment is to transfer knowledge of protein function from one species to another. Since sequence similarity metrics such as BLAST bit scores are not conclusive evidence of similar function, the purpose of aligning two PPI networks is to supplement sequence similarity with topological information so as to identify orthologs as accurately as possible. PIN alignment is relatively a young research area and successes of PIN network alignment so far include uncovering large shared sub-networks between species as diverse as *S. cerevisiae* and *H. sapiens*, and reconstructing phylogenetic relationships between species based solely on the amount of overlap discovered between their PPI networks [3]. Comparing two biological networks is a particularly challenging problem, since many interesting questions we might ask of these networks are computationally intractable to answer . Most papers in the literature report promising results in creating alignments that do indeed show large regions of biological or topological similarity between the PPI networks of various species, but few do both well [5, 6].

In this work, we try to align two or more PPI networks from different species. That is, we want to find a mapping from the nodes of one network to the nodes of another, in such a way as to maximize the topological and biological similarity of the pairs of nodes which are aligned to one another. This allows for the identification of orthologous proteins that are conserved during evolution as well as similar modules or pathways in the networks themselves.

We focus in particular on local network alignment. Existing approach for local network alignment fall in two main classes: (i) mine and merge approach, (ii) merge and mine approach. Algorithms of the first class at first analyze single networks, then integrate results. Conversely algorithms of the second class build an initial integrated network and then analyze such network [2, 7, **?**]. We propose MODULA, a novel local alignment method based on merge and mine approach. MODULA performs alignment using compact PIN modules or complexes extracted from two different species and explore best matching modules from them. Below we present the background of the study and few related works.

2 Problem Formulation and Related Work

Literature contains different formalizations of PIN alignment and we here follow the formalization we developed in a previous work by Mina and Guzzi [7].

Given two input graphs, $G_a = \{V_a, E_a\}$ and $G_b = \{V_b, E_b\}$, a correspondence between two regions of G_a and G_b can be expressed as a set of node pairs

$$S_i = \{(x^a, y^b) \mid x^a \in V_a \cup \eta,\ y^b \in V_b \cup \eta\} \tag{1}$$

where η is a fictitious symbol that means the associated protein has no ortholog in the other species.

Let

$$S_i^a = \{x^a \mid (x^a, y^b) \in S_i\}$$
$$S_i^b = \{y^b \mid (x^a, y^b) \in S_i\}$$

be the sets of proteins belonging to V_a and V_b, respectively, involved in S_i. Let G_i^a (G_i^b) be the subgraph induced by S_i^a (S_i^b) on G_a (G_b).

The **pairwise local network alignment problem** consists of finding all the correspondences S_i (i.e. groups of nodes) in order to maximize a cost function based on two criteria: (i) a similarity criterion that guarantees that matched subgraphs are topologically similar; and (ii) a model criterion drives the analysis toward the identification of specific topologies, and depends on the specific module to be uncovered (i.e. protein complex, linear pathway).

Literature contains many algorithms that have been proposed to detect conserved modules in PINs [8]. There exist different way to categorize such algorithms, we here distinct them on two main classes on the basis of the overall strategy: *mine and merge*, i.e. algorithms that first analyze each PIN separately, and then project solutions reciprocally from a PIN to the others [9, 10]; *merge and mine*, i.e. algorithms that fist integrate PINs into a single graph and then analyze such graph [11–13, 7].

Mine and merge analysis are usually less expensive in terms of computational resources, as evidenced by Erten et al. [9]. In general, merge and mine algorithms are more complicated due to difficulties in formulating and accounting for approximate matches, and the existence of multiple mappings between proteins in different species [9]. Moreover, they are computationally expensive, since in order to merge the input networks it is necessary to compare their topologies. The main drawback of these approaches is that these algorithms are more sensible to noise in input networks and to redundancy of information in input networks. Conversely, merge and mine algorithms are less sensible to these problems but they present in general a higher computational cost in the building of the initial integrated graph (also referred to as alignment graph). The interested reader may find a detailed discussion of these approaches in Mina and Guzzi[7].

A common problem in both groups is the requirement of additional information used to seed to build the alignment. Such seed are ortholog pairs and the absence of such information would require the exploration of all the possible combinations of protein pairs should be considered. Consequently many algorithms require as input two networks and a list of protein pairs, (for instance list of putative orthologs), to start the computation. List of pairs may be obtained from existing databases of orthologs, or gathering sequence similarity information using tools, using semantic similarity [14].

3 A New Alignment Approach

In this section we propose a new local alignment method MODULA that explore a matching between two different biologically significant compact protein inter-

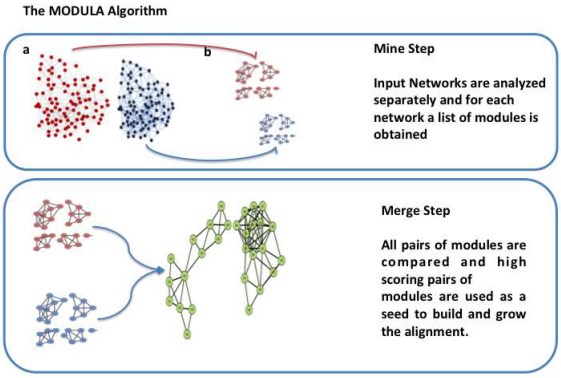

Fig. 1. Basic steps of MODULA.

action network modules from different species to find orthologous modules conserved functional similarity during evolution. Interestingly, the same approach may be extended for detecting conserved functional modules in multiple species. Broadly, MODULA is a two step process as describe below.

At first it identify compact network modules (or subgraphs) from input networks G_i and G_j using any state-of-the-art protein complex finding method. Then, every detected modules from G_i are compared and matched with each modules in G_j using any existing global alignment method.Finally, The best matching pairs are considered as aligned conserved modules.

The overall idea of the method is shown in Fig. 1. More formally, the step wise representation of MODULA is given in Algorithm 1.

3.1 The Algorithm

As an input other than two PINs from two different species, MODULA requires an user defined threshold τ to satisfy a minimum similarity score for global alignment between a pair of modules. To start with, MODULA needs compact modules or complexes. It uses existing network modules finding methods to detect biologically significant compact protein complexes (C_i and C_j). In our work, we use *ClusterOne* [15], an overlapping complex finding method from PINs. Recent study revels that ClusterOne is an effective technique in detecting biologically significant protein complexes [16]. For each module $M_i \in C_i$ is compared with each module $M_j \in C_j$ using any suitable global alignment method. We use here Magna++ for alignment [17]. MODULA considers only best match out of all pair matches. If the best match score is above threshold τ, it will be considered as best local alignment and added into the list L of all such alignments. The process continues for rest of the pairs.

Next, we assess the performance of our proposed method in light of several real data.

Algorithm 1: MODULA: The PIN Alignment Algorithm

Data: G_i (PIN-I); G_j (PIN-II); τ (Minimum Similarity Score)
Result: L (Aligned sub-networks)

$C_i \leftarrow$ `FindModules` (G_i);
$C_j \leftarrow$ `FindModules` (G_j);
// C_i and C_j list of compact modules detected by PIN complex finding method
for *each* $M_i \in C_i$ **do**
  for *each* $M_j \in C_j$ **do**
    if `Max` *(*`Alignment` *(*M_i, M_j *)) >* τ **then**
      L= L $\bigcup$ (M_i, M_j);
    end
  end
end
Return(L) ;

3.2 Experimental Evaluation

In order to assess performances of MODULA we compared it with respect to state of the art algorithms showing a sensible improvement of performances. In order to compare results, we expressed the performances of the algorithms in terms of the ability to recover known protein complexes conserved in two aligned species. Consequently, given a solution and a known complex, we measures this ability in terms of their overlap by using two classical measures: precision (π) and recall (ρ). Precision is defined as the fraction of proteins in the solution also present in the complex, while recall is the ration of proteins in the complex that are in common with the solution. Usually these measures are integrated into the F_1-score defined as the harmonic mean of precision and recall. Formally, these measures are defined as follows:

$$\pi = \frac{TP}{TP + FP},\ \rho = \frac{TP}{TP + FP},\ F_1 - score = \frac{2\pi\rho}{\pi + \rho}$$

where TP is the number of proteins found in a solution that are also in the complex. Analogously, FP and FN are the number of false positives and false negatives. The F_1-score ranges in the interval [0, 1], with 1 corresponding to perfect agreement. In our analysis, we match each known complex of a species to all the solutions of a given alignment, and we select as best match the solution with highest F_1-score. For each species we selected a dataset of known complexes as benchmark dataset. Within each dataset we identified many complexes with similar biological functions and highly overlapping with each other. This might lead to a biased evaluation since a solution might overlap with more than a known complex, and therefore be counted more than once. Moreover, these overlapping complexes are often quite small (3-4 proteins).

Comparison has been made against Align-MCL algorithms since it has been demonstrated AlignMCL outperformed other local alignment algorithms and also showed a more stability when different PINs of the same organisms are used. In order to compare MODULA with state-of-the-art methods, we use same datasets as used in the [13] comprises of interaction networks of mouse, yeast, human, worm and fly available in I2D database (release of 2011) [18]. The datasets

Table 1. Characteristics of Networks (Datasets) Used.

Species	Proteins	Interactions
D. Melanogaster [DM]	9854	37979
H. Sapiens [HS]	14567	138258
M. Musculus [MM]	4261	9547
C. Elegans [CE]	4755	9995
S. Cerevisiae [SC]	6182	147408

used here are presented in Table 1. However, for detail description of the dataset one may refer [13].

Initially, we clustered these networks using ClusterOne algorithm. Then for each alignment we built a comprehensive scoring matrix in which we compared all the pairs of generated modules. Finally, we used such pairs of high scoring modules to build a single aligned module. The resulting alignment is made by considering all the modules.

Table 2 summarizes results. Results shows that for this preliminary set of experiments, MODULA is able to recover more known complexes with respect to Align-MCL.

Table 2. Number of known complexes recovered by the different algorithms.

Alignment	Number of known complexes hit		Number of known complexes hit	
-	AlignMCL	MODULA	AlignMCL	MODULA
DM-SC	33	**35**	15	**16**
DM-HS	54	**55**	19	**25**
DM-CE	34	34	7	**9**
DM-MM	35	**42**	7	**16**

4 Conclusion

PPI networks are largely used to analyze biological mechanisms inside cells. Recently, many different experiments have generated a lot of data causing the growth of existing networks in terms of nodes and edges. Consequently, the need for the development of novel tools and methodologies for data management and analysis arose. In particular, one of the most exciting area is represented by the comparative analysis of protein interaction networks.

In this paper we proposed MODULA, a local network alignment algorithms that improves existing state of the art. The quality of the algorithm has been assessed. Results show that MODULA outperforms the other algorithms in discovering conserved functional modules (protein complexes).

A future work consists of comparing the solutions of different algorithms to determine their agreement. Additional assessments will be performed comparing the semantic similarity of the solutions.

References

1. M. Cannataro, P. H. Guzzi, and P. Veltri, "Protein-to-protein interactions," *ACM Computing Surveys*, vol. 43, no. 1, pp. 1–36, Nov. 2010.
2. ——, "Impreco: Distributed prediction of protein complexes," *Future Generation Computer Systems*, vol. 26, no. 3, pp. 434–440, 2010.
3. T. Milenković and N. Pržulj, "Topological characteristics of molecular networks," in *Functional Coherence of Molecular Networks in Bioinformatics*. Springer, 2012, pp. 15–48.
4. P. H. Guzzi, M. T. Di Martino, G. Tradigo, P. Veltri, P. Tassone, P. Tagliaferri, and M. Cannataro, "Automatic summarisation and annotation of microarray data," *Soft Computing*, vol. 15, no. 8, pp. 1505–1512, 2011.
5. P. H. Guzzi and T. Milenković, "Survey of local and global biological network alignment: the need to reconcile the two sides of the same coin," *Briefings in Bioinformatics*, p. bbw132, 2017.
6. M. T. Di Martino, P. H. Guzzi, D. Caracciolo, L. Agnelli, A. Neri, B. A. Walker, G. J. Morgan, M. Cannataro, P. Tassone, and P. Tagliaferri, "Integrated analysis of microrna, transcription factors and target genes expression discloses a specific molecular architecture of hyperdiploid multiple myeloma," *Oncotarget*, vol. 6, no. 22, pp. 19 132–47, 2015.
7. M. Mina and P. H. Guzzi, "Improving the robustness of local network alignment: Design and extensive assessmentof a markov clustering-based approach," *IEEE/ACM Trans. Comput. Biology Bioinform.*, vol. 11, no. 3, pp. 561–572, 2014. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TCBB.2014.2318707
8. J. Ji, A. Zhang, C. Liu, X. Quan, and Z. Liu, "Survey: Functional module detection from protein-protein interaction networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 99, no. PrePrints, p. 1, 2012.
9. S. Erten, X. Li, G. Bebek, J. Li, and M. Koyutürk, "Phylogenetic analysis of modularity in protein interaction networks." *BMC bioinformatics*, vol. 10, p. 333, Jan. 2009.
10. P. Jancura, E. Mavridou, E. Carrillo-de Santa Pau, and E. Marchiori, "A methodology for detecting the orthology signal in a PPI network at a functional complex level." *BMC bioinformatics*, vol. 13 Suppl 1, no. Suppl 10, p. S18, Jan. 2012.
11. G. Ciriello, M. Mina, P. H. Guzzi, M. Cannataro, and C. Guerra, "AlignNemo: A Local Network Alignment Method to Integrate Homology and Topology." *PloS one*, vol. 7, no. 6, p. e38107, Jan. 2012.
12. J. Flannick, A. Novak, C. B. Do, B. S. Srinivasan, and S. Batzoglou, "Automatic parameter learning for multiple local network alignment." *Journal of computational biology*, vol. 16, no. 8, pp. 1001–22, Aug. 2009.
13. M. Mina and P. H. Guzzi, "Alignmcl: Comparative analysis of protein interaction networks through markov clustering," in *2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops, BIBMW 2012, Philadelphia, USA, October 4-7, 2012*, 2012, pp. 174–181. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/BIBMW.2012.6470300

14. P. Guzzi, M. Mina, C. Guerra, and M. Cannataro, "Semantic similarity analysis of protein data: assessment with biological features and issues," *Briefings in bioinformatics*, vol. 13, no. 5, pp. 569–585, 2012.
15. T. Nepusz, H. Yu, and A. Paccanaro, "Detecting overlapping protein complexes in protein-protein interaction networks," *Nature methods*, vol. 9, no. 5, pp. 471–472, 2012.
16. P. Sharma, H. A. Ahmed, S. Roy, and D. K. Bhattacharyya, "Unsupervised methods for finding protein complexes from ppi networks," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 4, no. 1, pp. 1–15, 2015.
17. V. Vijayan, V. Saraph, and T. Milenković, "Magna++: Maximizing accuracy in global network alignment via both node and edge conservation," *Bioinformatics*, p. btv161, 2015.
18. K. R. Brown and I. Jurisica, "Unequal evolutionary conservation of human protein interactions in interologous networks," *Genome biology*, vol. 8, no. 5, p. R95, 2007.