A greedy and stochastic algorithm for multiple local alignment of interaction networks

G. Micale[1], A. Pulvirenti[2]✉, R. Giugno[2], A. Ferro[2]

[1]Department of Computer Science University of Pisa, Pisa, Italy

[2]Department of Clinical and Molecular Biomedicine, University of Catania, Catania, Italy

Motivations

A central problem in biological network analysis is the Local Network Alignment. The aim is to detect conserved subnetworks or complexes of proteins, across two or more species, which are involved in processes or functions. This allows to predict either new interactions or the functions of unknown proteins. Since the problem of finding conserved subnetworks in a set of networks is related to subgraph isomorphism, which is known to be NP-Hard, several heuristics have been proposed. These include PathBlast [1] and MaWISh [2] for pairwise local alignment and NetworkBlast-M [3] and Graemlin [4] for the multiple case. Although NetworkBlast-M has been proved to be the most efficient and accurate method, it is able to find significant conserved complexes composed by no more than 15 proteins. Here we introduce GASOLINE (Greedy And Stochastic algorithm for Optimal Local alignment of Interaction Networks) an algorithm based on Gibbs Sampling [5] in connection to a seed-extend approach to search for significant conserved complexes of any size.

Methods

The algorithm consists of two main phases. In the first phase, we look for ortholog proteins across the networks. We call these proteins seeds of the suboptimal pattern. In the second phase, called iterative phase, we extend each seed, by adding one adjacent node. Here, through a stochastic process based on Gibbs Sampling, we choose a node among a set of randomly picked adjacent ones. The chosen nodes will be those that maximize similarity among the N extended seeds. We repeat the iterative phase until we obtain a set of N conserved subgraphs each of size W. These N subgraphs represent our final alignment. The topological density and the conservation of a complex are measured through an Index of Density and Structural Conservation (hereafter IDSC). The IDSC score ranges from 0 to 1 and it is dynamically computed during the iterative phase. The iteration will terminate when IDSC is above a fixed threshold. This allows the removal of parameter W producing a set of highly dense and conserved complexes of different sizes.

Results

GASOLINE has been tested on 10 microbial PPI networks, taken from Graemlin [4] and on 6 eukaryotic PPI networks, taken from STRING database [6]. The number of proteins in microbial networks ranges from 1,000 to 7,000 with the amount of interactions ranging from 13,000 to 230,000, whereas the size of eukaryotic networks ranges from 6,000 to 12,500 proteins and from 26,000 to 166,000 edges. 2000 of execution of GASOLINE have been performed. As output we consider the best distinct (not overlapping) complexes, with respect to size andIDSC score. In our experiments we selected complexes of at least 5 proteins with IDSC score > 0.7. All tests have been performed on an Intel Core i5-2500 3.30Ghz CPU with 4 GB RAM.

The complexes computed by the algorithm have been then validated by annotating their proteins with GO categories. For the microbial networks, annotations have been downloaded from DAVID [7,8], while eukaryotic networks proteins have been annotated using BioDBNet [9]. Significant conserved categories have been obtained by computing a p-value (< 0.0001), based on hypergeometric distribution. A GO category has been considered conserved when it resulted significant in at least N-1 species, where N is the number of aligned networks.

The executions of GASOLINE on microbial networks revealed the existence of a big conserved complex of 40 proteins, forming the large and small subunits of ribosome, with IDSC equals to 0.755 (see Tab. 1). As for the 6 eukaryotic networks, 15 conserved complexes have been found by GASOLINE with IDSC greater than 0.75. They are listed in Tab. 2, with their IDSC and the number of GO categories enriched.

Significant GO categories	GASOLINE (Size = 40)	NetworkBlast-M (Size = 15)
GO:0003735 (structural constituent of ribosome)	1.887×10^{-16}	1.11×10^{-16}
GO:0005198 (structural molecule activity)	1.776×10^{-16}	9.992×10^{-17}
GO:0003723 (RNA binding)	1.665×10^{-16}	1.776×10^{-16}
GO:0019843 (rRNA binding)	1.443×10^{-16}	7.771×10^{-17}
GO:0006412 (translation)	8.882×10^{-17}	5.329×10^{-16}

Table. 1: Significant GO for microbial networks in GASOLINE and NetworkBlast-M

Complex	Size	IDSC	# GO categories enriched
Small and large subunit of ribosomes	43	0.716	7
Proteasome	32	0.847	10
Spliceosome	26	0.701	5
DNA-directed RNA polymerase	19	0.789	9
Small subunit (SSU) processome	15	0.832	2
Chaperonin-containing T-complex	13	0.737	4
V-ATPase	11	0.78	11
Exosome (RNase complex)	10	0.878	5
Replication fork protection	7	0.984	4
DNA replication factor C	7	0.928	5
Mitochondrial respiratory chain complex III	7	0.889	4
Arp2/3 protein complex	7	0.722	3
Translation initiation factor 2/2B	6	0.855	4
Cdc73/Paf1 complex	6	0.8	1
Endosomal sorting complex required for transport (ESCRT-III)	5	0.967	1

Table. 2: Conserved complexes found by GASOLINE in the 6 eukaryotic networks

References

1. Kelley BP, Sharan R, Karp R, Sittler ET, Root DE, Stockwell BR, and Ideker T. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc Natl Acad Sci U S A 100, 11394-9, 2003.
2. Koyuturk M, Kim Y, Topkara U, Subramaniam S, Szpankowski W, and Grama A, Pairwise alignment of protein interaction networks, Journal of Computational Biology, 13(2), 182-199, 2006
3. Kalaev M, Bafna V, and Sharan R. Fast and accurate alignment of multiple protein networks. Journal of computational biology, 16(8), 989–999, 2009.
4. Flannick J, Novak A, Srinivasan B, McAdams H, and Batzoglou S. Graemlin: general and robust alignment of multiple large interaction networks. Genome research, 16(9), 1169, 2006
5. Geman S, Geman D. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images". IEEE Transactions on Pattern Analysis and Machine Intelligence 6 (6): 721–741, 1984
6. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. 39 (Database issue):D561-8. Epub 2010 Nov 2, 2011.
7. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. 4(1):44-57, 2009.
8. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37(1):1-13, 2009
9. Mudunuri U, Che A, Yi M, and Stephens RM. BioDBnet: the biological database network. Bioinformatics. 25(4): 555–556, 2009