Abstracting Complex Interaction Networks

Lorenza Saitta♭, Corneliu Henegar♮, Jean-daniel Zucker♯♮
♭Università del Piemonte Orientale, Dipartimento di Informatica, Via Bellini 25/G, Alessandria, Italy;
♮UPMC Univ Paris 06, UMRS 872, NUTRIOMIQUE, CRC, 75006, Paris, France;
♯IRD, UMI 209, UMMISCO, IRD France Nord, F-93143, Bondy, France;

Abstract

The exploration of complex interaction networks has attracted considerable interest in various fields, ranging from fundamental biology and medicine to statistical physics and information technology. In "-omics" disciplines, significant progresses have been made in understanding the large-scale properties and the biological relevance of these interactions. Some properties such as "scale-free" distribution of nodes connectivity or "centrality" are aspects commonly described in such complex interaction systems. In many of these studies the analysis of network topology is complemented by a semantic analysis that may rely on different labels associated to the interacting entities. One of the bottleneck of these semantic analysis is that they are computationally costly. In this paper we present a framework to explore abstraction of networks useful to speedup the computation of ground network measures. Such abstraction mechanisms may be used to efficiently provide accurate approximations of ground network measures.

Introduction

The exploration of the complex molecular interactions defining the cellular environments is attracting considerable interest in various fields including biology and medicine. Supported by the unprecedented amount of biological sequence data generated by the Human Genome Project and by the development of *-omics* disciplines, significant progresses have been made in understanding the large-scale properties and the biological relevance of these interactions.

An important motivation for the study of *-omics* interactions resides in the ability of network formalisms to assess the biological relevance of a high number of interacting molecules in various experimental conditions. The interactional nature of the cellular processes, understood as associations of molecules whose relations to each other are instrumental in realizing a particular function (Alon 2003; Barabasi and Oltvai 2004; Hartwell et al. 1999), underlines the key role of the interaction patterns analysis in untangling the functional architecture of the cellular environments.

Copyright © 2009, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Uncovering patterns in large interaction networks is a difficult task because of the large number of nodes and edges. Thus, several proposals have been put forwards to extract manually or automatically useful information from these networks. To ease network exploration and analysis, Hu et al. advocate the use of multi-level approaches (Hu et al. 2007) which support zooming the network at different levels of granularity. Many other approaches rely on clustering nodes (Huang, Wei, and Pan 2006; Huang and Pan 2006) based on their topological properties and/or their associated labels. Corneliu et al., for example, use node labels and make them diffuse on the graph to generate clusters that are then used to target useful biological information (Henegar et al. 2006). Unfortunately the computation that are required at the ground level is classically super-quadratic with the size of the network. In this extended abstract we define a framework for network abstractions, so as to explore the feasibility for computing approximations of typical ground statistics based on the abstract network level.

Biological Problem Description

The strong relation between the biological roles of molecules and the modular organization of their interactions has been long hypothesized even before high-throughput genomic data became available (Hartwell et al. 1999). Several studies have uncovered correlations of network centrality indices (be it *connectivity* or *betweenness* between nodes in the network) with indicators of biological relevance, such as lethal knockout phenotypes (Jeong et al. 2001). Other studies have shown that specific patterns of phylogenetic variability (Guimera and Nunes Amaral 2005) were also correlated with centrality indices.

In parallel, the biological interpretation of high-throughput gene expression measurements (*i.e.*, the genomic functional profiling), has evolved into a highly standardized analytical framework. Functional profiling tools rely on curated biological annotation resources, as those provided by the Gene Ontology Consortium (GO) (Ashburner et al. 2000) or the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al. 2008), and on statistical significance measures to

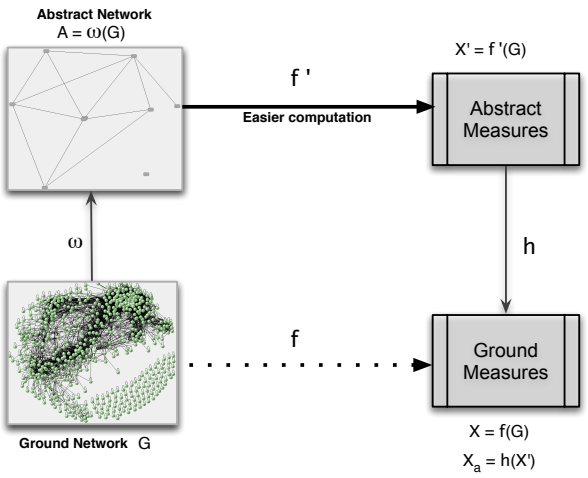

Figure 1: Schema of the abstraction mechanism to compute abstract measures using an abstract network

identify biological themes, which are significantly over-represented among those annotating a set of transcripts. These tools became popular, as they provide valuable insights into the biological phenomena encoded into microarray expression measurements.

Although these two aspects of the biological interpretation of transcriptomic signatures have been developed independently, they follow a common goal by exploring complementary facets of the biological phenomena encoded into gene expression profiles. The benefit of integrating them into a unique framework is suggested by the strength of the relation linking the modular structuring of transcripts interactions to their biological roles, as well as by the interactional nature of the latter ones, which cannot be conceived independently from those of the other molecular actors that compose the cellular environments.

Algorithmic solutions inspired by network formalisms are of major interest for integrating heterogeneous information pertaining to the analysis of genomic data. Among recent developments, several proposals relied on network abstractions to represent semantic relations between GO biological themes annotating gene products and to improve the characterization of the functional profile of gene expression datasets (Aubry et al. 2006; Chabalier, Mosser, and Burgun 2007). Another goal is to increase the accuracy of supervised and unsupervised classification methods available for microarray data by integrating information on relevant molecular interactions extracted from pathways annotation databases such as KEGG. Finally, several authors proposed various ways of incorporating the biological information available on transcripts roles (*i.e.* the sharing of common annotating themes) to weight similarity metrics and improve the biological relevance of the clustering of transcriptional profiles (Zhou, Kao, and Wong 2002; Huang, Wei, and Pan 2006; Henegar et al. 2006).

A recent approach has suggested to take into account the biological knowledge available on transcripts roles to improve the biological relevance of the modular patterns identified in co-expression networks (Prifti et al. 2008; Henegar et al. 2008). Such an approach explores both transcriptional and functional patterns in co-expression networks, and it has proven to be more biologically relevant in terms of transcriptional and functional patterns identified in co-expression networks.

Network Abstraction Operators

The proposed approach is summarized in Figure 1. Starting from a ground interaction graph $\mathcal{G}$, we are interested in computing some measurements on it, for instance the centrality of every node or its betweenness. Let $X = f(\mathcal{G})$ be the set of such measures, obtained by means of the procedures globally denotes f. For large networks, these computations may be too long, or their results may remain hidden in the large amount of ground information, making difficult understanding their meaning. As an alternative, we generate an abstract network $\mathcal{A} = \omega(\mathcal{G})$ through the application of an abstraction operator ω, which "condenses" the ground network. Then, we extract from $\mathcal{A}$, by means of procedures f', the properties that are analogous of those looked for in $\mathcal{G}$, namely $X' = f'(\mathcal{A})$. Finally, we transform the values of properties X' into the values of properties X again, by applying the functions h and obtaining $X_a = h(X')$. In this process, the requirement that $X_a = X$ may be too strong, and we accept that X_a be only an approximation of X.

More formally, let $\mathcal{G}$ be a ground undirected network with n vertices (nodes), each one associated to a specific gene in a given set. In $\mathcal{G}$ two nodes g_i and g_j are connected with an arc if they are co-expressed, *i.e.*, they show a similar expression pattern. More precisely, co-expression between two genes can be quantified by assigning to the connecting arc a *weight*, namely a positive real number w_{ij} that represents its strength. With this representation, the ground network is a *weighted graph*. In order to work with *unweighted graphs*, a thresholding process can be applied to the arcs: given a threshold τ, all arcs whose associated weights are less than τ are removed. The unweighted graph is a special case of the weighted one, occurring when all weights are either 0 (no connecting arc) or 1 (connecting arc). In both cases, the ground graph is represented through its adjacency matrix $\mathcal{M}_g$. For the sake of exemplification, in Figure 2 a gene matrix $\mathcal{G}$ with $n = 574$ gene is reported. Let us consider now a set $A = \{a_1, ..., a_m\}$ of *themes*, *i.e.*, labels annotating genes and referring to some particular biological role. For instance, themes may be *biological processes*, in which the gene is involved, or *molecular component*, in which the gene is present, and so on. The labels corresponding to the themes are attached to the nodes of graph $\mathcal{G}$. In analogy with the ground space, a network $\mathcal{A}$ can be associated to the set A. The nodes of $\mathcal{A}$ correspond to themes, and the arcs represent similarity between themes. Usually, the network

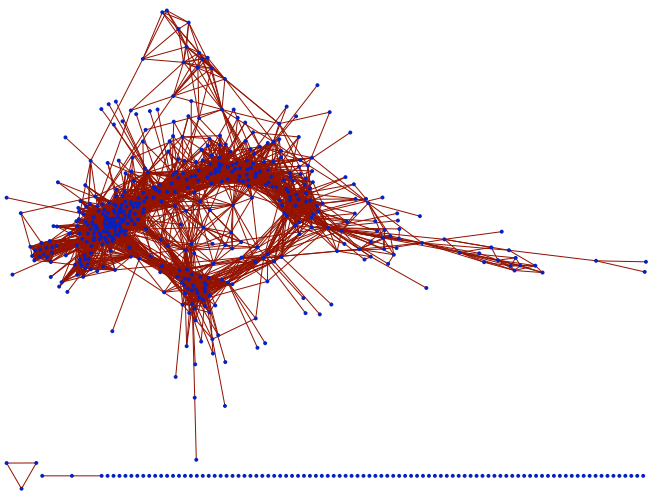

Figure 2: Example of a gene matrix $\mathcal{G}$ with $n = 574$ nodes (genes).

$\mathcal{A}$ is a weighted one and it is very dense. Graph $\mathcal{A}$ is represented by its adjacency $m \times m$ matrix $\mathcal{M}_a$. Referring to Figure 1, we can formally say that graph $\mathcal{A}$ is derived from $\mathcal{G}$ by means of the abstraction operator ω:

$$\mathcal{A} = \omega(\mathcal{G}) \tag{1}$$

Equation (1) is a synthetic way of expressing the fact that the matrix $\mathcal{M}_a$ can be derived from $\mathcal{M}_g$ through the procedure Ω, associated to the formal operator ω:

$$\mathcal{M}_a = \Omega(\mathcal{M}_g) \tag{2}$$

It is convenient, in the envisaged application, to consider Ω as associated to an $n \times m$ matrix $\mathcal{W}$, which establishes the links between genes (corresponding to the rows) and themes (corresponding to the columns). An entry $w_{i,r}$ $(1 \leq i \leq n, 1 \leq r \leq m)$ in $\mathcal{W}$ represents the strength with which gene g_i is associated to theme a_r. The procedure Ω is composed of two steps: in the first one, it fills the matrix $\mathcal{M}_a$, and, in second one, computes pairwise distances between the columns (themes) of $\mathcal{M}_a$, determining thus the edges of $\mathcal{A}$. By changing either the way matrix $\mathcal{W}$ is filled or the distance measure between themes, different abstraction operators can be implemented.

By summarizing, $\mathcal{M}_g$ represents the co-expression similarity between genes, matrix $\mathcal{M}_a$ represents the similarity between themes, and matrix $\mathcal{W}$ represents the association between subsets of genes and themes.

Experimentations

The most basic type of abstraction that can be thought of is the association of identically labelled genes to the abstract node corresponding to their label. An entry in the corresponding matrix, denoted $\mathcal{W}^{(b)}$, will be $w^{(b)}_{i,r} = 1$, if gene g_i has label a_r, and $w^{(b)}_{i,r} = 0$ otherwise. In $\mathcal{W}^{(b)}$ each column represents the set of genes "covered" by a theme, and it can be thought of as a feature vector describing the theme. By applying a distance measure to these feature vectors, we can construct the abstract basic matrix $\mathcal{A}^{(b)}$, whose entries contain pairwise similarity between themes. In order to treat the theme on a uniform basis, each column is normalized, by dividing each entry by the sum of the corresponding column. In this basic case, the similarity between two themes is determined by the number of genes that bear both labels. Euclidean distance is used afterwards.

A more complex way of constructing $\mathcal{W}$, that is derived from biological consideration (Henegar et al. 2006), treats a theme label as a "fluid" that originates from the genes covered by the theme and spreads into the ground network through the nearest neighbors of these last and the nearest neighbors of nearest neighbors and so on. More precisely, let a consider an initial, normalized matrix $\mathcal{W}^{(0)} = \mathcal{W}^{(b)}$, and let us iterate the computation of $\mathcal{W}^{(t)}(t \geq 1)$ starting from $\mathcal{W}^{(t-1)}$ and $\mathcal{M}_g$, and renormalizing the columns at each iteration step:

$$\mathcal{W}^{(t)}_{temp} = \mathcal{M}_g \; \mathcal{W}^{(t-1)} \tag{3}$$

$$w^{(t)}_{i,r} = \frac{w^{(t)}_{i,r}}{\sum_{k=1}^{n} w^{(t)}_{i,r,temp}} \tag{4}$$

By applying Equation (3) four times, convergence is reached to a final matrix $W^{(\infty)}$. In Figure 3 the abstract graph $\mathcal{A}^{(\infty)}$ corresponding to the ground network of Figure 2 is reported. Given the two graphs, we have experimented the comparison sketched in Figure 1 for the parameters: *closeness centrality*, *betweenness*, *clustering index* and *degree distribution*. In the following, only the results for the centrality c are reported, for the sake of brevity. Let $\vec{x}$ be the vector (of size n) containing the values of the centrality computed directly on the nodes of $\mathcal{G}$. Let $\vec{x'}$ be the vector (of size m) containing the values of the centrality computed on the nodes of $\mathcal{A}$. We compute the approximation of the ground values of the centrality for a ground node (gene) as the sum of the centralities of the abstract themes that cover it, weighted by the strength of the link between the gene and the theme:

$$c(g_j) = \sum_{r=1}^{m} \left(c(a_r) \; \frac{w^{(\infty)}_{j,r}}{\sum_{k=1}^{m} w^{(\infty)}_{j,k}} \right) \quad (1 \leq j \leq n) \tag{5}$$

We have compared the centrality computed directly on the ground graph and the one derived from the abstract graph. The results are reported in Figure 4. As we can see, the average behavior on the whole set of nodes is almost the same, even though there are differences on the single nodes. For the betweenness the correspondence is less marked, whereas for the clustering index the results are similar to Figure 4. The degree distribution is a power law on both the ground and the abstract network even though with different exponents.

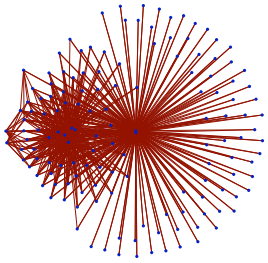

Figure 3: Abstract matrix $\mathcal{A}^{(\infty)}$ obtained from equation (4). The nodes (themes) are $m = 145$.

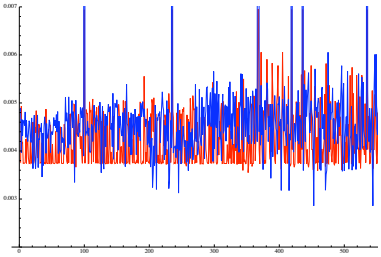

Figure 4: Comparison between the centrality computed directly on the ground graph (blue) and the one derived from the abstract graph (red). On the abscissa there are the nodes' identifiers. The two graphs have been scaled to obtain superposition. As it is apparent, the average behaviors of the two measures agree, and, what is more important, both measures agree on the high peaks. This means that the used abstraction is able to capture the highest centrality nodes.

Conclusion

In this paper we have reported an initial work on using abstraction to simplify the analysis of complex networks, in particular biological networks. The idea is to reduce the computational complexity of the analysis and to increase the interpretability of the results by abstracting the ground network into a much smaller one. For this initial investigation the abstraction operator has been suggested by biological considerations, but we intend, in the future, to try to learn the best ones that satisfy a set of given contraints.

References

Alon, U. 2003. Biological networks: the tinkerer as an engineer. *Science* 301(5641):1866–7.

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.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; and Sherlock, G. 2000. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet* 25(1):25–9.

Aubry, M.; Monnier, A.; Chicault, C.; de Tayrac, M.; Galibert, M. D.; Burgun, A.; and Mosser, J. 2006. Combining evidence, biomedical literature and statistical dependence: new insights for functional annotation of gene sets. *BMC Bioinformatics* 7:241.

Barabasi, A. L., and Oltvai, Z. N. 2004. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 5(2):101–13.

Chabalier, J.; Mosser, J.; and Burgun, A. 2007. A transversal approach to predict gene product networks from ontology-based similarity. *BMC Bioinformatics* 8:235.

Guimera, R., and Nunes Amaral, L. A. 2005. Functional cartography of complex metabolic networks. *Nature* 433(7028):895–900.

Hartwell, L. H.; Hopfield, J. J.; Leibler, S.; and Murray, A. W. 1999. From molecular to modular cell biology. *Nature* 402(6761 Suppl):C47–52.

Henegar, C.; Cancello, R.; Rome, S.; Vidal, H.; Clement, K.; and Zucker, J. D. 2006. Clustering biological annotations and gene expression data to identify putatively co-regulated biological processes. *J Bioinform Comput Biol* 4(4):833–52.

Henegar, C.; Tordjman, J.; Achard, V.; Lacasa, D.; Cremer, I.; Guerre-Millo, M.; Poitou, C.; Basdevant, A.; Stich, V.; Viguerie, N.; Langin, D.; Bedossa, P.; Zucker, J.; and Clement, K. 2008. Adipose tissue transcriptomic signature highlights the pathological relevance of extracellular matrix in human obesity. *Genome Biology* 9(1):R14.

Hu, Z.; Mellor, J.; Wu, J.; Kanehisa, M.; Stuart, J. M.; and DeLisi, C. 2007. Towards zoomable multidimensional maps of the cell. *Nat Biotech* 25(5):547–554.

Huang, D., and Pan, W. 2006. Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data. *Bioinformatics* 22(10):1259–68.

Huang, D.; Wei, P.; and Pan, W. 2006. Combining gene annotations and gene expression data in model-based clustering: weighted method. *OMICS* 10(1):28–39.

Jeong, H.; Mason, S. P.; Barabasi, A. L.; and Oltvai, Z. N. 2001. Lethality and centrality in protein networks. *Nature* 411(6833):41–2.

Kanehisa, M.; Araki, M.; Goto, S.; Hattori, M.; Hirakawa, M.; Itoh, M.; Katayama, T.; Kawashima, S.; Okuda, S.; Tokimatsu, T.; and Yamanishi, Y. 2008. Kegg for linking genomes to life and the environment. *Nucleic Acids Res* 36(Database issue):D480–4.

Prifti, E.; Zucker, J. D.; Clement, K.; and Henegar, C. 2008. Funnet: an integrative tool for exploring transcriptional interactions. *Bioinformatics* 24(22):2636–8.

Zhou, X.; Kao, M. C.; and Wong, W. H. 2002. Transitive functional annotation by shortest-path analysis of gene expression data. *Proc Natl Acad Sci U S A* 99(20):12783–8.