BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 24 2010, pages 3083–3089
doi:10.1093/bioinformatics/btq591

Gene expression

Advance Access publication October 19, 2010

Interactional and functional centrality in transcriptional co-expression networks

Edi Prifti[1,2,3,*], Jean-Daniel Zucker[1,2,3], Karine Clément[1,2,4] and Corneliu Henegar[1,2,4,*]

[1]INSERM, UMR-S 872, Les Cordeliers, Eq. 7 Nutriomique, Paris, F-75006, [2]Pierre et Marie Curie-Paris 6 University, Cordeliers Research Center, UMR-S 872, Paris F-75006, [3]IRD, UMI 209, UMMISCO, IRD France Nord, Bondy F-93143 and [4]Assistance Publique – Hôpitaux de Paris, Pitié Salpêtrière Hospital, Nutrition department, Paris F-75013, France

Associate Editor: Trey Ideker

ABSTRACT

Motivation: The noisy nature of transcriptomic data hinders the biological relevance of conventional network centrality measures, often used to select gene candidates in co-expression networks. Therefore, new tools and methods are required to improve the prediction of mechanistically important transcriptional targets.

Results: We propose an original network centrality measure, called annotation transcriptional centrality (ATC) computed by integrating gene expression profiles from microarray experiments with biological knowledge extracted from public genomic databases. ATC computation algorithm delimits representative functional domains in the co-expression network and then relies on this information to find key nodes that modulate propagation of functional influences within the network. We demonstrate ATC ability to predict important genes in several experimental models and provide improved biological relevance over conventional topological network centrality measures.

Availability: ATC computational routine is implemented in a publicly available tool named FunNet (www.funnet.info)

Contact: edi.prifti@crc.jussieu.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 15, 2010; revised on September 24, 2010; accepted on October 15, 2010

1 INTRODUCTION

In complex cellular environments, biological functions are emerging properties that cannot be understood without taking into consideration the system as a whole. Molecular interactions allow for cellular components to function and for cells to live and perform their tasks within the organism. It is through chemical and physical interactions that molecules influence one another and carry out biological functions. Since some of these influences may have a greater impact than others, a hierarchical characterization of their relative importance can be very useful in exploring the functional architecture of the cellular environments. The concept of interactional centrality is borrowed from social sciences where it is used to estimate the relative importance of individuals within societies (Freeman, 1977). Biological processes, viewed as associations of molecules whose relations to each other assure a particular function (Barabasi and Oltvai, 2004; Hartwell *et al.*, 1999), have a strong interactional nature. Understanding interaction patterns is therefore crucial in untangling the functional architecture of cellular environments.

Network analysis plays an increasingly important role in the exploration of complex interactional systems in numerous domains. In biology, they are used to study regulation of the transcriptome, proteome, metabolome, methylome and other cell systems (Kepes, 2007). Key properties of biological networks are shared with other complex interaction systems, such as the '*scale-free*' distribution of node connectivity (Barabasi and Oltvai, 2004) or the '*small-world*' aspect of their fast synchronization (Watts and Strogatz, 1998). Others, like the existence of a hierarchical architecture of modules formed by functionally related components (i.e. genes, enzymes and metabolites) (Ravasz *et al.*, 2002), or the presence of various types of '*hubs*' (i.e. highly connected or central nodes), which modulate key interactions at distinct levels in the cell, are thought to have a particular relevance in biology. The existence of strong relationships between the biological roles of molecules and the modular organization of their interactions was demonstrated through analyses of integrating gene expression patterns with information on phylogenetic variability, sharing of common transcriptional binding sites, results from mutagenesis experiments and available knowledge on genes' biological roles (Allocco *et al.*, 2004; Bergmann *et al.*, 2004; Carlson *et al.*, 2006; Guimera and Nunes Amaral, 2005; Hartwell *et al.*, 1999; Jeong *et al.*, 2001).

Hub molecules with important biological properties are identified in these networks through topological interactional centrality measures. Relying on results from mutagenesis experiments, Jeong *et al.* (2001) showed that highly connected hub proteins that occupy central positions in the network architecture of yeast protein–protein interactions are three times more likely to be biologically essential than those with only a small number of links in the network. Several other studies have shown that proteins with high betweenness centrality scores (i.e. measure of total number of shortest paths that connect each two nodes in the network passing through a given one) are more likely to be functionally essential (Joy *et al.*, 2005; Yu *et al.*, 2007). Others have demonstrated strong correlations of network topological centrality indices with specific patterns of phylogenetic variability (Guimera and Nunes Amaral, 2005).

Microarray technology has been very successful in biomedical research over the past decade. The study of gene expression

*To whom correspondence should be addressed.

profiles in functional genomics and drug discovery is one of the most important applications of microarrays. The analysis of co-expression patterns using network representations is widely used for exploring the large amount of data produced by this high-throughput technology. Also called co-expression networks, these abstractions illustrate the complex relationships between expression profiles of individual genes. They are built by relating transcripts (i.e. the nodes of the network) that display similar expression profiles (i.e. the edges connecting the nodes) following conventional analytical frameworks (Zhang and Horvath, 2005). Topological centrality measures computed in network models, such as node degree and betweenness, are used to identify molecular targets in well-characterized protein–protein or regulation networks. In transcriptional co-expression networks, the high degree of experimental noise limits, however, the biological relevance and robustness of centrality measures, based on topological criteria alone, in predicting important transcripts (Wu, 2009). Integrating expression data with available biological knowledge of genes may help increase the biological pertinence of such measures. We describe here an integrative approach that relies on a two-layer network model, built by adding a functional layer on top of the transcriptional one based on functional annotations extracted from genomic databases. This approach allows delimiting functional domains in the co-expression network, and then uses this information to identify key nodes that play important roles in the propagation and modulation of functional influences. To this purpose, an original centrality measure called *annotation transcriptional centrality* (ATC) was designed to quantify the influence of each transcriptional node upon the propagation of functional themes in the co-expression network. We demonstrate that in the context of co-expression networks, ATC is more significantly effective than conventional topological indicators in predicting functionally important transcripts.

$$\begin{array}{c|cccc} & A_1 & A_2 & \dots & A_m \\ \hline t_1 & W_{1,1} & W_{1,2} & \dots & W_{1,m} \\ t_2 & W_{2,1} & W_{2,2} & \dots & W_{2,m} \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & \dots & \mathrm{M} \\ t_n & W_{n,1} & W_{n,2} & \dots & W_{n,m} \end{array}$$

Fig. 1. The initial configuration T^0 of the non-linear dynamical system in which the transcripts $\{t_1,\dots,t_i,\dots,t_n\}$, represented in the rows, are considered as instances of their annotating themes $\{A_1,\dots,A_j,\dots,A_m\}$, appearing in the columns, with $W_{i,j}^p = 1$ if $t_i \subset A_j$, and $W_{i,j}^p = 0$ otherwise.

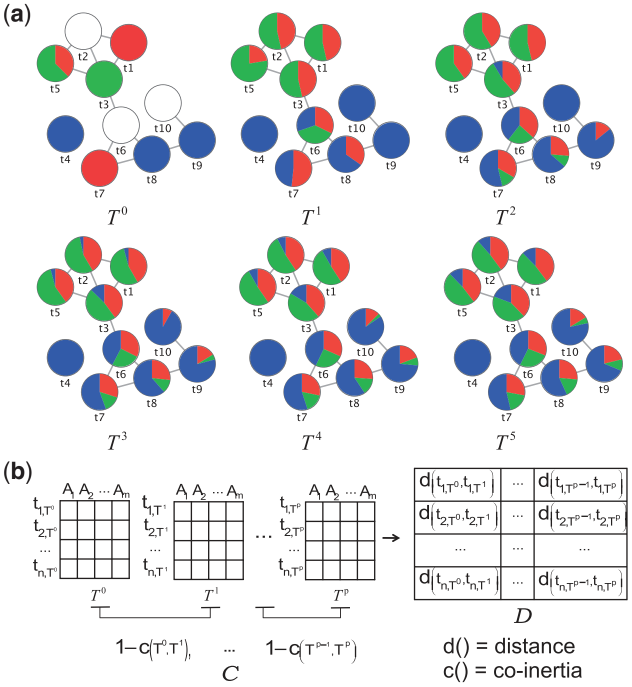

Fig. 2. (**a**) Propagation of the genomic annotations within an example network composed of 10 transcript nodes. Three annotations represented by the colors blue, green and red propagate within the network until convergence is reached after five steps. Node t_4 does not change its annotation during the process since it is not connected to the others. (**b**) The different configuration matrices $\{T^0,T^1,\dots,T^p\}$ are used to compute ATC scores.

2 METHODS

2.1 ATC

The knowledge gathered on functional roles of genes and proteins has been structured and made available in public databases such as Gene Ontology (GO) (Ashburner *et al*., 2000) and KEGG (Kanehisa and Goto, 2000). GO is organized around three main axes: (i) biological process, (ii) molecular function and (iii) cellular component. ATC computation algorithm starts by identifying genomic annotations from public databases, which are significantly overrepresented in the analyzed expression data. Transcripts are related to their corresponding annotations, represented as sets of transcriptional instances, to build a two-layer network model by adding a functional layer on the top of the transcriptional co-expression network (Henegar *et al*., 2008; Prifti *et al*., 2008) (see also Supplementary Section 3 for more information on the functional annotation procedure). A non-linear dynamical system is then used to simulate the propagation of these genomic annotations within the transcriptional co-expression network and compute a functional measure of interactional centrality for each transcript. The initial configuration of the dynamic system is generated by assigning weights to the categorical values of a relational table that represents the annotation setup (i.e. transcripts and their genomic annotations), illustrated in Figures 1 and 2. The strategy employed for updating intermediary configurations of the system relies on a transcriptional adjacency matrix representing the co-expression network, as well as on a combining operator '⊕'. The edges relating transcriptional nodes in the co-expression network can be valued either in a weighted (i.e. in the interval [0,1], using a soft thresholding technique) or an unweighted manner (i.e. as 1 or 0 for similar or non-similar, through a hard thresholding approach) (Zhang and Horvath, 2005). We used *addition* as combining operator because of its straightforward interpretation and the well-understood convergence properties of the associated non-linear dynamical systems (Gibson *et al*., 2000; Zhang *et al*., 2000). Considering this operator, the amount of weight $W_{i,j}^p$ of a theme A_j associated with a transcript-node t_i in a configuration T^p (Figs 1 and 2) is computed as the sum of the weights of A_j available in the previous configuration T^{p-1} in all transcript-nodes connected to t_i (i.e. the dot product between the column vector A_j^{p-1} and the row vector t_i of the adjacency matrix computed from transcriptional expression measurements). The column vectors of T^p are normalized to the unit after each step in order to keep constant the total amount of weight associated with each annotating theme. The dynamical system convergence is assessed using a combined criterion associating a co-inertia analysis with a Mantel test. Co-inertia is a multivariate approach that identifies trends or co-relationships in complex datasets, and is particularly useful when the number of variables exceeds the number of observations. It is therefore well suited for a large number of analytical situations involving microarray data (Culhane *et al.*, 2003). The dynamical system is considered convergent when the co-inertia of two consecutive configurations is greater than a threshold of 0.9 (i.e. on a [0,1] scale) and they demonstrate a significant similarity assessed with a Mantel test (Supplementary Fig. S2). The propagation of functional attributes within the co-expression network is schematized in Figure 2a and Supplementary Figure S1 where three annotations (represented by colors blue, green and red) propagate in a 10-node network.

After the system's convergence the role of each node in the modulation and propagation of functional attributes within the co-expression network is quantified by considering: (i) local information showing the amount of weight associated with each annotation that passes through each node between each two consecutive configurations and (ii) global information indicating how much the system as a whole changes after each iteration. The amount of weight associated with all annotations of a given node is computed using the Euclidean distance between the annotation vectors $t_{i,T^{p-1}}$ and t_{i,T^p} of each two consecutive configurations and the results are stored in the distance matrix D, as shown by Equation (1) and Figure 2b.

$$D_{i,p} = \sqrt{\sum_{i=1}^{n} \left(t_{i,T^p} - t_{i,T^{p-1}}\right)^2} \qquad (1)$$

This matrix is further weighted by multiplying each column vector $D_{[1,n],p}$ with the global distance between each two consecutive configurations, computed in relation to their co-inertia $C_p = 1 - \text{co-inertia}(T^{p-1}, T^p)$ (vector C in Fig. 2b). Eventually, the rows of the resulting weighted matrix D' are summed to compute the vector of ATC values. The algorithm was written in R language for statistical computations (R Development Core Team, 2010) and was implemented into the FunNet package starting with version 1.08 (see Supplementary Section 5 for the pseudo code of the dynamical system and the computation of ATC). It is also available via a web tool at http://www.funnet.info (Prifti *et al.*, 2008).

2.2 Experimental assessment

To evaluate the usefulness of ATC in spotting key functional transcripts in co-expression networks, we analyzed its relationships to empirical indicators of functional essentiality, derived from mutagenesis experiments, and computational indicators of sequence phylogenetic conservation in yeast. We used public microarray data related to cell-cycle conditions, since cell cycle involves both phylogenetically conserved biological processes, such as metabolism and transportation, as well as less conserved ones, such as transcriptional regulation and signaling (Lopez-Bigas *et al.*, 2008). Another biological indicator of relevance in relation to cell cycle is transcriptional periodicity, a measure of the relation between gene expression and cell-cycle phases. Data from mutagenesis experiments were provided by *Saccharomyces* genome deletion project (Winzeler *et al.*, 1999). Our choice in selecting these biological indicators was based on previous evidence from the literature, which demonstrate that biologically essential proteins, associated with non-viable phenotypes when mutated, are very connected and central in protein–protein networks (Jeong *et al.*, 2001; Yu *et al.*, 2007). They are also highly conserved phylogenetically (Drummond *et al.*, 2005; Jordan *et al.*, 2002) due to a number of mechanisms that operate to restrain the variability of genomic sequences involved in key cellular processes. Eventually, robustness to the experimental noise affecting microarray data of ATC and topological centrality measures was comparatively evaluated. This assessment was performed by incorporating increasing amounts of random noise in co-expression networks built from yeast and human microarray data.

2.2.1 Phylogenetic conservation score We defined a phylogenetic conservation score inspired by existing work on assessing phylogenetic conservation of genomic sequences (Lopez-Bigas *et al.*, 2008). Each gene sequence $sp_{ref,i}$ of the reference species sp_{ref} (i.e. *Saccharomyces Cerevisiae* in our case) was aligned against the complete genomes of nine other species sp_j: *Caenorhabditis Elegans*, *Arabidopsis Thaliana*, *Drosophila Melanogaster*, *Gallus Gallus*, *Danio Rerio*, *Strongylocentrotus Purpuratus*, *Mus Musculus*, *Canis Familiaris* and *Homo Sapiens* using discontiguous megablast with a 16 bit long coding template and a word size of 12 (Ma *et al.*, 2002). Only the best alignment scores $s_{ref,i,j}$ with E-values <1 were considered for each species (Fig. 3a). Since the alignment score depends on the sequence length, all similarity scores were normalized by dividing them with the similarity scores of the sequences aligned to themselves (2) as proposed by Lopez-Bigas *et al.* (2008). Considering the fact that phylogenetic distances between species are variable, we used them to weight the conservation scores (3). The phylogenetic distance between the reference species and the nine others is condition relative since it is computed in relation to a limited number of genes and not to the entire genomes.

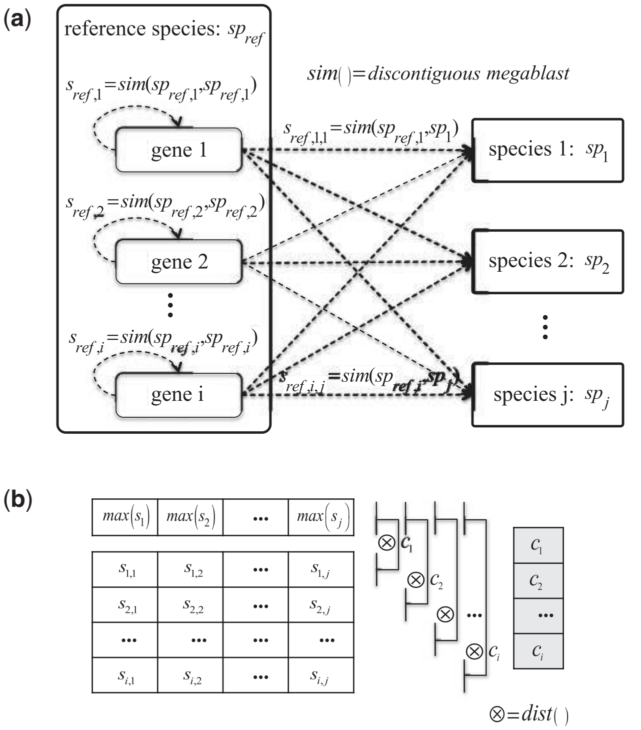
(a)
reference species: sp_{ref}
$sim(\,)$=discontiguous megablast
$s_{ref,1}=sim(sp_{ref,1},sp_{ref,1})$
gene 1
$s_{ref,2}=sim(sp_{ref,2},sp_{ref,2})$
gene 2
$s_{ref,i}=sim(sp_{ref,i},sp_{ref,i})$
gene i
$s_{ref,1,1}=sim(sp_{ref,1},sp_1)$
$s_{ref,i,j}=sim(sp_{ref,i},sp_j)$
species 1: sp_1
species 2: sp_2
species j: sp_j
(b)
$max(s_1)$ $max(s_2)$ ••• $max(s_j)$
$s_{1,1}$ $s_{1,2}$ ••• $s_{1,j}$
$s_{2,1}$ $s_{2,2}$ ••• $s_{2,j}$
••• ••• ••• •••
$s_{i,1}$ $s_{i,2}$ ••• $s_{i,j}$
c_1
c_2
•••
c_i
$\otimes = dist(\,)$

Fig. 3. (**a**) Schematic illustration of the alignment of i genes from the studied species and the genomes of other species. (**b**) Matrix with normalized alignment scores $s_{i,j}$ of all studied genes $sp_{ref,i}$ of the reference species sp_{ref} (on the rows) against the other species sp_j (on the columns). The unified conservation score c_i is computed as the Euclidean distance between the vector of alignment scores for each gene $(s_{i,1}, s_{i,2}, \ldots, s_{i,j})$ and the vector containing the highest alignment scores for each species $(\max(s_1), \max(s_2), \ldots, \max(s_j))$.

$$s'_{ref,i,j} = \frac{\text{sim}\left(sp_{ref,i}, sp_j\right)}{\text{sim}\left(sp_{ref,i}, sp_{ref,i}\right)} \qquad (2)$$

$$s_{i,j} = s'_{ref,i,j} \times \left(1 - \text{dist}\left(sp_{ref}, sp_j\right)\right) \qquad (3)$$

Similarity scores in relation to the different species were further used to compute a unified conservation score c_i as the Euclidean distance between the vector of normalized similarity scores $(s_{i,1}, s_{i,2}, \ldots, s_{i,j})$ for each gene in all species and the vector containing the highest similarity scores for each species $(\max(s_1), \max(s_2), \ldots, \max(s_j))$ (Fig. 3b). This unified conservation score ranges from 0, when no significant similar sequence is found in the selected species, to 1 when the very same sequence is found in all the other species. Genomic datasets as well as the alignment programs were downloaded from NCBI ftp servers and installed locally. The selection of the different organisms was conditioned by the relatively small number of completely sequenced genomes.

2.2.2 Microarray expression data Microarray expression data from six different yeast cell-cycle experiments with a total of 123 cDNA-chips were used in this study (Cho *et al*., 1998; de Lichtenberg *et al*., 2005; Pramila *et al*., 2006; Spellman *et al*., 1998). These data were downloaded from Cyclebase (Gauthier *et al*., 2008), a centralized database where the authors normalized, synchronized and assigned periodicity significance *P*-values illustrating the relation between gene expression and the cell-cycle phases. Only the expression profiles of 410 genes, which showed significant periodicity in all six experiments, were selected to build transcriptional co-expression network used in subsequent analyses (see Supplementary Section 3). The robustness of the various centrality measures to the experimental noise affecting microarray data was evaluated based on yeast expression data, as well as on a human dataset obtained from a series of white adipose tissue biopsies performed in a set of 25 morbidly obese and 10 lean subjects (Henegar *et al*., 2008).

2.2.3 Biological relevance of ATC Saccharomyces Genome Deletion Project knocked off 90% of yeast's genome and revealed that 17% of the genes are essential for the cell to survive (Winzeler *et al*., 1999). We relied on these data to compare ATC scores of essential and non-essential genes. Non-parametric Wilcoxon rank tests were performed to assess whether essential genes (74 out of 410) had significantly higher ATC scores than non-essential ones. ATC scores were computed using annotations from three different databases: GO Biological Process (GOBP), GO Cellular Component (GOCC) and KEGG.

Next, we computed the phylogenetic conservation scores for the 410 periodic genes, as explained above, and tested whether genes with the highest phylogenetic conservation had higher ATC scores than less-conserved ones. Since the phylogenetic conservation scores are continuous values, we used the first and last quartiles of the conservation score distribution to label genes (i.e. conserved and non-conserved genes). Wilcoxon rank test was computed to compare ATC scores of the two gene categories.

The ability of ATC in predicting biologically important genes linked to cell cycle was also tested in relation to the transcriptional periodicity score proposed in Cyclebase (Gauthier *et al*., 2008) (see also Supplementary Section 3), based on the assumption that highly periodic genes are the most functionally important for the cell cycle. Similarly, as for the phylogenetic conservation scores, genes from the first quartile of the periodicity score distribution (i.e. those with the smallest periodicity *P*-values) were labeled as periodic and those from the last quartile as non-periodic. Wilcoxon rank tests were computed to compare ATC scores of the two gene categories in this situation, as well as to test for additional relationships between phylogenetic conservation, transcriptional periodicity and gene essentiality.

Zotenko *et al.* (2008) showed that in yeast protein interaction networks, essential proteins tend to cluster in densely connected subnetworks with other proteins that are involved in the same biological processes, and that these subnetworks are rich in hubs. We investigated whether this was also the case for the transcriptional co-expression network of yeast cell cycle. A spectral clustering approach (Ng *et al*., 2001) was used in association with Silhouette partition quality criteria to identify transcriptional modules within the network. Fisher's exact tests were further performed to evaluate the enrichment of transcriptional modules in essential genes (i.e. to test whether essential genes tend to cluster together in the same module).

2.2.4 Comparative assessment of topological centrality measures and ATC Topological centrality measures, such as degree and node betweenness centrality, were shown to be effective for identifying essential molecules in well-characterized interaction networks such as yeast protein–protein interaction or regulation networks (Jeong *et al*., 2001; Yu *et al*., 2007). The same methodological approach as the one used in Section 2.2.3 for ATC was employed to assess the ability of degree and node betweenness centralities to predict essential, phylogenetically conserved or periodic genes. The usefulness of ATC and topological centrality measures in predicting essential genes in co-expression networks was compared through receiver operating characteristic (ROC) curves analysis (DeLong *et al*., 1988).

The robustness to noise of the three centrality measures was analyzed by adding increasing levels of uniform random noise (1, 10, 20, 30, 40 and 50% of the $n(n-1)$ number of possible edges composing of an n-node network) to the co-expression network computed form the original data through the hard thresholding approach described in Section 2.1. For each noise level, we performed 30 consecutive iterations and computed the mean values of each of the three centrality measures for every transcript. We then used non-parametric Wilcoxon tests to assess whether the original centrality values computed without noise, expressed as percentages of the total sum of nodes' centrality scores in the network, were significantly different from those computed after adding increasing levels of noise.

3 RESULTS

3.1 ATC relations with biological indicators of relevance

Our assessment shows that essential genes have significantly higher ATC scores than non-essential ones ($P < \times 10^{-5}$, Fig. 4a), irrespective of the annotation systems used to compute these scores. Figure 5a shows in red essential genes with ATC values in the first quartile of the distribution computed with GOCC and in blue essential genes outside the first quartile of the distribution (48.6% of essential genes are found in the first quartile of GOCC ATC score distribution).

The analysis of the phylogenetic sequence conservation, another indicator of functional importance, demonstrated an equally strong relationship with ATC regardless of which annotation systems were used to compute these scores. Figure 5b shows that 51.2% of the genes belonging to the first quartile of the phylogenetic conservation score distribution (i.e. red nodes) are also found in the first quartile of ATC score distribution. Subsequent tests showed that genes with the highest phylogenetic conservation scores (i.e. the first quartile of the distribution) also have higher ATC values than those with lower conservation scores (i.e. the last quartile of the distribution; $P < 5 \times 10^{-10}$, Fig. 4b).

Furthermore, ATC showed a significant ability to predict contextual functional importance, represented in our case by the transcriptional periodicity during cell-cycle phases. Our evaluation showed that genes with lower periodicity *P*-values, and thus higher periodicity ranks, have higher ATC values than less periodic genes ($P < 5 \times 10^{-20}$, Fig. 4c), with 54.9% of periodic genes found in the first quartile of ATC distribution (Fig. 5c, nodes in red).Additional analysis showed that essential genes display strong phylogenetic conservation of their sequences compared with non-essential genes ($p < 8 \times 10^{-12}$). This result highlights the biological relevance of the computational score of sequence phylogenetic conservation, in agreement with previous work (Guimera and Nunes Amaral, 2005). Moreover, highly conserved genes among the 410 displaying significant transcriptional periodicity during cell cycle (i.e. with sequence conservation scores in the first quartile of the distribution), demonstrated a significantly higher transcriptional periodicity than less-conserved genes ($P < 2 \times 10^{-16}$). Finally, spectral clustering combined with Silhouette optimization criterion identified two transcriptional modules in the yeast cell-cycle co-expression network, with a majority of the essential genes clustering together in one of the modules ($P < 0.05$). The

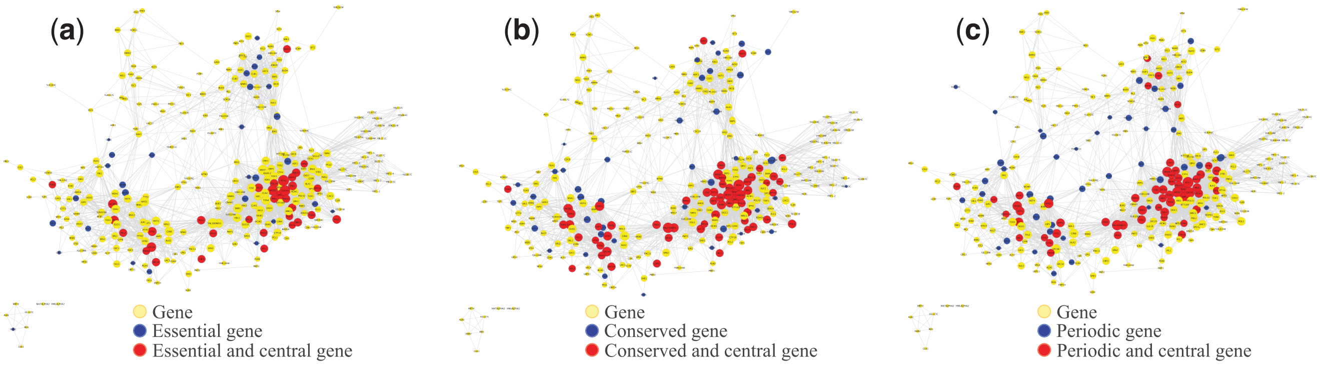
(a)
● Gene
● Essential gene
● Essential and central gene
(b)
● Gene
● Conserved gene
● Conserved and central gene
(c)
● Gene
● Periodic gene
● Periodic and central gene

Fig. 5. Transcriptional co-expression network of yeast periodic genes: (**a**) essential genes with high ATC scores are depicted in red while those with low ATC scores are depicted in blue and non-essential genes in yellow; (**b**) genes with high phylogenetic conservation and high ATC scores (i.e. the first quartile of the respective score distributions) are depicted in red, while those with high conservation scores but low ATC values are indicated in blue and those with low conservation scores in yellow; (**c**) genes with high transcriptional periodicity and high ATC scores (i.e. the first quartile of the respective score distributions) are depicted in red, while those with high transcriptional periodicity but low ATC values are indicated in blue and those with low transcriptional periodicity in yellow. The size of the nodes reflects their ATC scores computed with GOCC annotation system.

functional characterization of these modules showed distinct functional profiles (see Supplementary Fig. S4). Module 1 regroups transcripts mostly involved in nuclear processes related to DNA metabolism (i.e. '*DNA metabolic process*',*'chromatin assembly*','*DNA recombination*','*DNA repair*','*nucleus*','*nuclear chromosome part*', '*nuclear chromatin*',etc.), while module 2 is mainly annotated by themes associated with cytoplasmic processes in relation to cell division (i.e. '*interphase*','*microtubule-based movement*','*cytoskeletal part*','*microtubule part*','*cytoplasmic microtubule*','*cell division site part', etc.*).

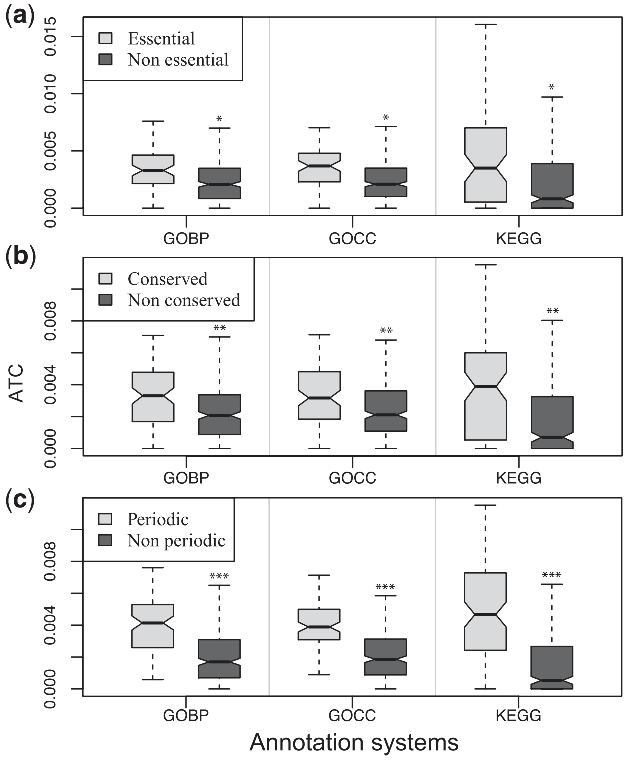

Fig. 4. Boxplots with ATC values computed for three annotation systems GOBP, GOCC and KEGG are plotted for (**a**) essential versus non-essential, (**b**) phylogenetically conserved versus non-conserved and (**c**) periodic versus non-periodic genes. The differences between these groups are statistically significant (*$P<5\times10^{-5}$, **$P<5\times10^{-10}$, ***$P<5\times10^{-20}$).

3.2 Comparative assessment of ATC and topological centrality measures

The individual assessment of the topological centrality measures (degree and node betweenness centrality) in relation with the mentioned indicators of biological relevance, demonstrated that essential genes have higher degree centrality scores than non-essential ones ($P<0.05$), but similar node betweenness scores. This result supports previous work showing higher connectivity of essential proteins in protein–protein interaction networks (Jeong *et al.*, 2001).

Furthermore, genes with high phylogenetic conservation scores also demonstrated significantly higher degree and betweenness scores than less phylogenetically conserved ones ($P<5\times10^{-15}$). A similar direct association was observed between transcriptional periodicity scores and the two topological centrality measures ($P<5\times10^{-30}$).

Comparative assessment through ROC analysis of ATC and topological centralities showed significantly better abilities of ATC, computed either with GOCC or KEGG, to identify essential genes ($P<0.05$, Fig. 6a).

Finally, no significant difference between the original ATC scores and those computed after adding noise was observed, even for a 50% level of random noise (see Supplementary Figs S5 and S6 in Section 4 of the Supplementary Material). In contrast, similar assessments performed for each of the two topological centrality measures demonstrated a very significant sensitivity to the random noise, starting from the lowest noise level of 1%, particularly for network betweenness that displayed the highest noise sensitivity.

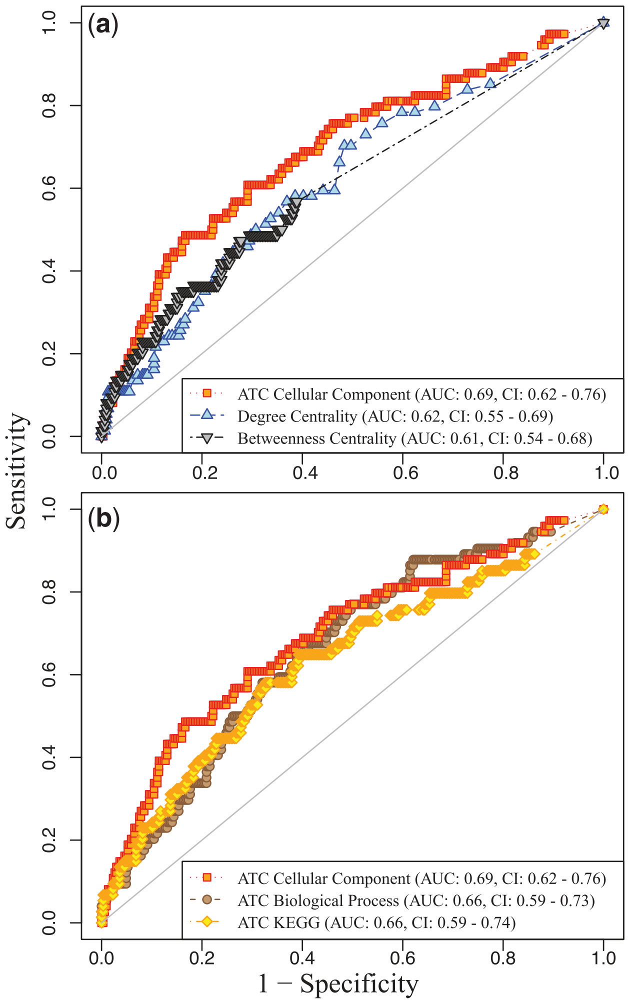

Fig. 6. (**a**) ROC analysis comparing the abilities of ATC (computed with GOCC annotations), degree and node betweenness centralities in predicting essential genes; (**b**) ROC analysis comparing the abilities of ATC computed with GOCC, GOBP and KEGG annotation systems in predicting essential genes. The area under the curve and the 95% confidence intervals are denoted between parentheses.

4 DISCUSSION

Our assessment shows that conventional topological centrality indexes are not equally useful in predicting essential genes in yeast cell-cycle co-expression networks. Indeed, degree centrality appeared to perform better than betweenness in spotting essential genes. This finding contradicts previous data published by Yu *et al.* (2007), which showed that node betweenness centrality is a better indicator of functional essentiality than degree in regulatory and other directed networks. However, as suggested by Zotenko *et al.* (2008), in yeast protein interaction networks essential proteins tend to participate in biological processes that are densely interconnected, and which are therefore more likely to be hubs (i.e. highly connected nodes). In our context, the advantage of degree centrality over betweenness could be explained by an increased sensitivity of the later to the high amount of noise, which is often wired in the structure of the co-expression networks. This observation is also supported by our assessment, which associates the highest sensitivity to artificial random noise with betweenness centrality (see Supplementary Tables S1 and S2). On the other hand, ATC predicts essentiality better than both betweenness and degree centrality in the transcriptional co-expression network built from cell-cycle data. Our assessment indicated also that ATC is significantly more robust to experimental noise than topological measures. This advantage may be related to the fact that ATC computation relies not only on the topology of the co-expression network but also on its functional architecture.

Furthermore, no significant difference between ATC distributions computed with the three annotation systems (GOCC, GOBP and KEGG) was observed (Fig. 6b). This finding suggests an equal usefulness of the available annotation systems, despite the fragmentary character of the background knowledge on which they rely. It is expected that annotation coverage and its biological precision will continue to increase in the future and thus further improve ATC prediction abilities.

5 CONCLUSIONS

In this study, we proposed an original centrality measure, the ATC, aiming to improve the prediction of biologically relevant gene targets within transcriptional co-expression networks. ATC computation relies on the integration of microarray expression profiles with biological knowledge extracted from public genomic databases, which enrich the biological signal in transcriptomic data. Its predictive abilities were assessed in relation to conventional centrality measures based on topological criteria alone. It would be interesting to further test ATC in application to other 'omic' networks, such as protein–protein interaction, regulation and metabolic networks. From a different perspective, the biological signal encoded in co-expression networks can be further enriched by integrating other types of data, pertaining to gene transcription (i.e. DNA methylation), to improve the biological relevance of interactional centrality measures.

ACKNOWLEDGEMENTS

This work was promoted by the 'Direction de la Recherche Clinique (DRC)/Assistance Publique Hôpitaux de Paris'. We wish to thank the three anonymous referees for their helpful comments that allowed us to improve this work.

Funding: INSERM (PRNH No. 4NU10G); the 'Agence Nationale de la Recherche' (RIOMA program No. ANR-05-PCOD-030-02); the European Community 6th framework HEPADIP project; grant provided by the 'Association de Langue Française pour l'Etude du Diabète et des Maladies Métaboliques' (ALFEDIAM to C.H.).

Conflict of Interest: none declared.

REFERENCES

Allocco,D.J. *et al*. (2004) Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics*, **5**, 18.

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

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

Bergmann,S. *et al*. (2004) Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol.*, **2**, E9.

Carlson,M.R. *et al*. (2006) Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics*, **7**, 40.

Cho,R.J. *et al*. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.

Culhane,A.C. *et al*. (2003) Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics*, **4**, 59.

de Lichtenberg,U. *et al*. (2005) New weakly expressed cell cycle-regulated genes in yeast. *Yeast*, **22**, 1191–1201.

DeLong,E.R. *et al*. (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, **44**, 837–845.

Drummond,D.A. *et al*. (2005) Why highly expressed proteins evolve slowly. *Proc. Natl Acad. Sci. USA*, **102**, 14338–14343.

Freeman,L.C. (1977) A set of measures of centrality based on betweenness. *Sociometry*, **40**, 35–40.

Gauthier,N.P. *et al*. (2008) Cyclebase.org-a comprehensive multi-organism online database of cell-cycle experiments. *Nucleic Acids Res.*, **36**, D854–D859.

Gibson,D. *et al*. (2000) Clustering categorical data: an approach based on dynamical systems. *VLDB J.*, **8**, 222–236.

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

Hartwell,L.H. *et al*. (1999) From molecular to modular cell biology. *Nature*, **402**, C47–C52.

Henegar,C. *et al*. (2008) Adipose tissue transcriptomic signature highlights the pathological relevance of extracellular matrix in human obesity. *Genome Biol.*, **9**, R14.

Jeong,H. *et al*. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Jordan,I.K. *et al*. (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.*, **12**, 962–968.

Joy,M.P. *et al*. (2005) High-betweenness proteins in the yeast protein interaction network. *J. Biomed. Biotechnol.*, **2005**, 96–103.

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

Kepes,F. (2007) *Biological Networks*. World Scientific Publishing Company, New Jersey.

Lopez-Bigas,N. *et al*. (2008) Functional protein divergence in the evolution of Homo sapiens. *Genome Biol.*, **9**, R33.

Ma,B. *et al*. (2002) PatternHunter: faster and more sensitive homology search. *Bioinformatics*, **18**, 440–445.

Ng,A. *et al*. (2001) On spectral clustering: analysis and an algorithm. In Dietterich,T. *et al*. (eds) *Advances in Neural Information Processing Systems 14*. MIT Press, Cambridge, USA, pp. 849–856.

Pramila,T. *et al*. (2006) The Forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle. *Genes Dev.*, **20**, 2266–2278.

Prifti,E. *et al*. (2008) FunNet: an integrative tool for exploring transcriptional interactions. *Bioinformatics*, **24**, 2636–2638.

R Development Core Team. (2010) *R Foundation for Statistical Computing*. Vienna, Austria; 2010. R: A Language and Environment for Statistical Computing. ISBN 3-900051-07-0.

Ravasz,E. *et al*. (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.

Spellman,P.T. *et al*. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Watts,D.J. and Strogatz,S.H. (1998) Collective dynamics of ‘small-world’ networks. *Nature*, **393**, 440–442.

Winzeler,E.A. *et al*. (1999) Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science*, **285**, 901–906.

Wu,Z. (2009) A review of statistical methods for preprocessing oligonucleotide microarrays. *Stat. Methods Med. Res.*, **18**, 533–541.

Yu,H. *et al*. (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput. Biol.*, **3**, e59.

Zhang,B. and Horvath,S. (2005) A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.*, **4**, Article17.

Zhang,Y. *et al*. (2000) Clustering categorical data. In *Proceedings of the 16th International Conference on Data Engineering*, San Diego, California, USA, p. 305.

Zotenko, E., Mestre, J., O’Leary, D.P. and Przytycka, T.M. (2008) Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput Biol.*, **4**, e1000140.