BIOINFORMATICS ORIGINAL PAPER
Vol. 27 no. 22 2011, pages 3149–3157
doi:10.1093/bioinformatics/btr546

Systems biology
Advance Access publication September 29, 2011

# TFRank: network-based prioritization of regulatory associations underlying transcriptional responses

Joana P. Gonçalves[1,2,*], Alexandre P. Francisco[1,2], Nuno P. Mira[3,4], Miguel C. Teixeira[3,4], Isabel Sá-Correia[3,4], Arlindo L. Oliveira[1,2] and Sara C. Madeira[1,2,*]

[1]Knowledge Discovery and Bioinformatics group, INESC-ID, [2]Department of Computer Science and Engineering, Instituto Superior Técnico (IST), Technical University of Lisbon, [3]Institute for Biotechnology and Bioengineering, IST, Technical University of Lisbon and [4]Department of Bioengineering, IST, Technical University of Lisbon, Lisbon, Portugal

Associate Editor: Alfonso Valencia

## ABSTRACT

**Motivation:** Uncovering mechanisms underlying gene expression control is crucial to understand complex cellular responses. Studies in gene regulation often aim to identify regulatory players involved in a biological process of interest, either transcription factors coregulating a set of target genes or genes eventually controlled by a set of regulators. These are frequently prioritized with respect to a context-specific relevance score. Current approaches rely on relevance measures accounting exclusively for direct transcription factor–target interactions, namely overrepresentation of binding sites or target ratios. Gene regulation has, however, intricate behavior with overlapping, indirect effect that should not be neglected. In addition, the rapid accumulation of regulatory data already enables the prediction of large-scale networks suitable for higher level exploration by methods based on graph theory. A paradigm shift is thus emerging, where isolated and constrained analyses will likely be replaced by whole-network, systemic-aware strategies.

**Results:** We present TFRank, a graph-based framework to prioritize regulatory players involved in transcriptional responses within the regulatory network of an organism, whereby every regulatory path containing genes of interest is explored and incorporated into the analysis. TFRank selected important regulators of yeast adaptation to stress induced by quinine and acetic acid, which were missed by a direct effect approach. Notably, they reportedly confer resistance toward the chemicals. In a preliminary study in human, TFRank unveiled regulators involved in breast tumor growth and metastasis when applied to genes whose expression signatures correlated with short interval to metastasis.

**Availability:** Prototype at http://kdbio.inesc-id.pt/software/tfrank/.

**Contact:** jpg@kdbio.inesc-id.pt; sara.madeira@ist.utl.pt

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

Received on March 23, 2011; revised on August 31, 2011; accepted on September 25, 2011

## 1 INTRODUCTION

Transcription is the first step in the flow of biological information from genome to proteome and its tight regulation serves as a crucial checkpoint in most biological processes. Transcriptional control often results from the concerted action of diverse transcription factors (TFs), acting alone or as interacting partners, regulating or being targeted, and participating in complex feedback loops (Lemon and Tijian, 2000; Teixeira *et al.*, 2010). Genome-wide studies have been unveiling intricate regulatory networks with increasing layers of complexity as the hierarchy is amplified, empowered by the rapid accumulation of regulatory data in public repositories. Namely, YEASTRACT (Abdulrehman *et al.*, 2011) contains TF–target links for *Saccharomyces cerevisiae*, while JASPAR (Sandelin *et al.*, 2004) and TRANSFAC (Wingender, 1996) collect binding sites and position-specific scoring matrices for diverse eukaryotes including human. These data provide information on the topology of the regulatory network, which is static and thus not sufficient *per se* to determine whether a given association is relevant or not in a specific biological context. In fact, binding of a given TF to the promoter region of its target(s) is rather an intermittent event triggered by biochemical stimuli and ultimately leading to changes in expression. Insight on the mechanisms underlying the stimulus–response effect can only be achieved through an integrated study of control and behavior, where three challenges often arise: (i) identification of regulators of genes involved in a similar, related biological process; (ii) identification of genes coregulated by a group of TFs of interest; and (iii) prioritization of regulators, target genes or regulatory interactions with respect to a measure of relevance expressing their contribution within the regulatory network under study.

Diverse strategies have been proposed to address the first challenge. A popular one is to rank TFs based on the ratio of genes targeted in the dataset under analysis. This score can be normalized by the relative number of targets in the genome to avoid excessive bias toward TFs controlling very large regulons (Teixeira *et al.*, 2006). Alternatively, methods may rely on a statistical framework to discover enriched target binding sites by calculating their overrepresentation in the test sequences against a background model estimated from a collection of (randomly) selected sequences (Hestand *et al.*, 2008; Marstrand *et al.*, 2008; Veerla *et al.*, 2010), eventually considering cross-species conservation (Hestand *et al.*, 2008; Veerla *et al.*, 2010). A drawback of such procedure is that results are highly susceptible to the choice of the background model and variations on the statistical parameters. Other strategies have overcome this limitation by computing the genome-wide overrepresentation of binding sites (Chang *et al.*, 2006; Ho Sui *et al.*,

*To whom correspondence should be addressed.

2005; Roider *et al.*, 2009), improved through comparative analysis with orthologous sequences based on phylogenetic footprinting. Regulatory associations have also been uncovered from expression data alone (Pournara and Wernisch, 2007; Reverter *et al.*, 2010) by searching for correlation of expression profiles under the assumption that TFs tend to express coherently with their targets. However, this is often not the case. In addition, these techniques tend to ignore physical binding and thus generate a larger number of false positives (Balleza *et al.*, 2009; Margolin and Califano, 2007). The issue was recently mitigated by incorporating binding, expression data and three types of control: regulation, activation and inhibition (Essaghir *et al.*, 2010). Most methods for the second challenge are experimental, relying on ChIP-chip and DNA microarray assays to find putative targets. Computationally, the problem has been addressed by computing the genome-wide enrichment of binding sites (Chang *et al.*, 2006), or the correlation between the expression of each gene with the composite expression of sets of genes pertaining identical binding sites (Kim and Kim, 2006).

All revised approaches include some form of prioritization based on a relevance score and most are able to test large-scale data. Nonetheless, they restrict to direct neighbors in the network, disregarding overlapping control and systemic effects of utter importance in gene regulation, where cascade responses proliferate. This systemic perspective was supported in recent work by Booth *et al.* (2010), suggesting that evolution might lead to the introduction of new layers of control between TFs and their targets. Embedding this knowledge into the relevance score would likely increase informativeness and accuracy in determining key regulatory players, particularly when coupled with the ability to direct the search toward wider-purposed or more specific regulators (Martínez-Antonio, 2003; Vallet-Gely *et al.*, 2007). Another characteristic of the methods reviewed herein is that they are mostly sequence driven and thus not tailored for the study of higher level hypothesis on gene regulation. Notably, their ability to predict TF–target affinities can be exploited in a first step to derive regulations and build large-scale regulatory networks for methods based on graph theory, known to deliver effective and efficient network analysis. Notably, network motif mining has been successfully applied to the identification of regulatory modules (Bar-Joseph *et al.*, 2003; Zhang *et al.*, 2008). This technique aims to partition the network in coherent groups and is thus more appropriate for finding subgroups of activity rather than studying the composite effect generating a particular response.

In this article, we present TFRank, a method to prioritize regulatory players, involved in a process of interest, considering all known regulatory paths of a given organism. The most common use case is to rank the TFs controlling a set of genes responsive under specific conditions, according to a score calculated as follows. Each gene in the network is given an initial weight indicative of its presence (or absence) in (from) the gene set. Expression values or any alternative measure of importance may be used. A personalized ranking strategy is then applied to propagate the weights through the network graph in the target→TF direction. For the prioritization of targets of a set of regulators, the diffusion is performed in the TF→target direction instead. TFRank is independent from network construction and may be applied to any organism with available TF–target interactions. TFRank provides: (i) full topology analysis including alternative regulatory paths with intermediate layers of control; (ii) ability to regulate preference for more global or local players; (iii) integrated stimulus–response analysis of (static) binding and (dynamic) response; (iv) potential incorporation of interaction confidence scores; and (v) normalization to mitigate bias toward TFs with a large number of targets or genes acted upon by a large number of regulators. We evaluate the contribution of each TFRank parameter to the ranking on real data. We test the effectiveness of TFRank to unravel regulators potentially involved in the transcriptional control of biological processes in yeast and human. For yeast, we examine the TFs output by TFRank for two sets of genes upregulated in response to stress induced by the antimalarial drug quinine and the food preservative acetic acid. For human, we present a preliminary study of the TFRank-suggested regulators of genes whose expression signature has been previously associated with short interval to metastasis in breast cancer.

## 2 METHODS

This section introduces TFRank, a framework to: (i) prioritize TFs of a set of genes responsive under particular conditions; and (ii) rank the genes controlled by a group of regulators involved in a biological process of interest (Fig. 1). Application of the method to one setting or the other is straightforward and depends exclusively on the representation devised for the regulatory network, whether using reverse directed regulatory interactions from target genes to regulators (first problem) or the original regulations from TFs to their targets (second problem). Our description focuses mainly on the solution for the first problem, more frequently addressed. Answering the reverse question should also become clear from the following exposition. We first outline the definitions and formulate the problem. Next, we introduce TFRank's personalized ranking strategy based on the heat kernel. Finally, we describe the contribution of each parameter.

### 2.1 TFRank prioritization framework

A regulatory network is defined as a directed graph $G=(V,E)$, where $V$ denotes the set of vertices, comprising regulators and target genes, and $E$ is the set of edges representing regulatory associations between elements in $V$. Let $A$ and $D$ denote the adjacency and diagonal matrices of $G$, respectively, where $A_{uv}=w(u,v)$ expresses the weight of the edge between a source vertex $u$ and an end vertex $v$ and $D_{uu}=d(u)$ is the outdegree of vertex $u$ (or the sum of weights of the outgoing edges). Note that, for the first problem, the direction of each regulatory interaction is reversed, meaning that edge sources correspond to target genes and end vertices represent regulators. This is equivalent to using the transpose of the original regulatory network graph. Given a set $T\subseteq V$ of target genes, the first problem aims at defining a relevance score and thus obtaining a ranking on $R\subseteq V$, where $R$ is the set of TFs regulating directly or indirectly the target genes in $T$.

Personalized PageRank (Brin and Page, 1998) is a widely known solution for the above formulation, related to local clustering on graphs (Andersen and Lang, 2006; Chung, 2007). Personalized PageRank involves a preference vector $p_0$, which can be regarded as the probabilistic distribution of the genes in $T$, and a jumping constant $\alpha$ denoting the probability of returning back to initial nodes. For an input $p_0$, the ranking $p_\alpha$ is given by the following equation and equivalent recurrence, respectively,

$$p_\alpha=\alpha\sum_{k=0}^{\infty}(1-\alpha)^k p_0(D^{-1}A)^k \tag{1}$$

$$p_\alpha=\alpha p_0+(1-\alpha)p_\alpha D^{-1}A. \tag{2}$$

This is also equivalent to the lazy definition of PageRank, using matrix $M=\frac{1}{2}\left(I+D^{-1}A\right)$ instead of $D^{-1}A$, up to a change in $\alpha$.

TFRank uses the heat kernel rank, which was shown recently to perform better (Chung, 2007) than PageRank. The heat kernel is a fundamental solution for the heat equation describing the diffusion of heat and related temperature variation through a bounded region over time. In physics, the

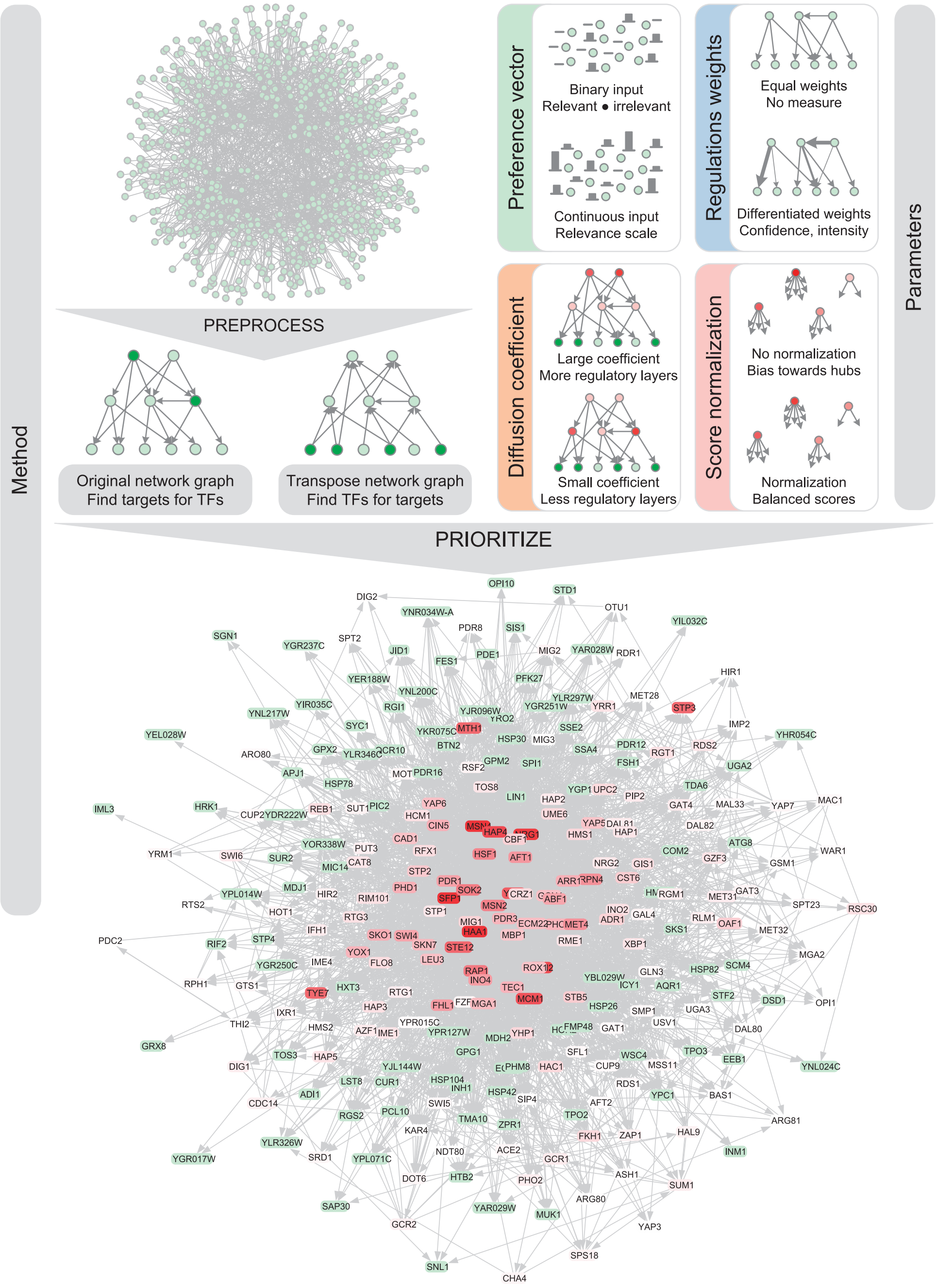

**Fig. 1.** Diagram of the TFRank framework for prioritizing TF–target interactions. TFRank uses the whole regulatory network and a preference vector accepting two kinds of input: a list of genes, containing for example genes whose expression was significantly altered in a given condition; or a list of genes together with their expression levels or any measure of importance. Based on the input, each gene is initialized with a given score as follows: in the first case, the score can be either 1 or 0 depending on whether the gene is in the input list or not; in the second case, the score is set as the absolute value of the given measure. These scores, expressed in the form of a preference vector, are then propagated through the network to compute a TFRank relevance score for every regulatory player, involving two additional properties: regulation weights and diffusion coefficient. Regulation weights can be used to express a degree of confidence or intensity on each regulatory interaction, taking into consideration the nature of the underlying experiments. The diffusion coefficient allows to control the extent of intermediate regulatory layers considered. Finally, genes are ranked based on the computed TFRank scores. Normalization can be applied to balance the final scores and avoid bias toward regulators targeting many genes or genes acted upon by many TFs. In this figure, nodes are generally represented in light green (standard node or target) or red (regulator). Dark green is used to denote a gene present in the input list. The bottom figure highlights a ranking obtained with TFRank using YEASTRACT regulations (Abdulrehman *et al.*, 2011). The intensity of the red nodes indicates the relevance of the corresponding TF upon prioritization: the darker the color, the more relevant the regulator. Prefuse (Heer *et al.*, 2005) was used to generate the bottom network from custom code.

diffusion is defined by a partial differential equation of the form

$$\frac{\partial}{\partial t} p_t = -p_t(I-W), \tag{3}$$

where $t$ is the thermal conductivity or heat diffusion coefficient and $W = D^{-1}A$. The solution of the heat equation is given by an exponential kernel (the heat kernel), defined in Equation (4), where $p_0$ is again a preference vector.

$$p_t = e^{-t} \sum_{k=0}^{\infty} \frac{t^k}{k!} p_0 W^k \tag{4}$$

Analogously, in graph theory, $W$ denotes the transition probability matrix of a typical random walk on $G$ defined as $W = D^{-1}A$. In this case, $W$ holds the weights of TF–target interactions between pairs $(u,v)$ normalized by the sum of weights of the outgoing edges of $u$. Other matrices can be used, such as the Laplacian $D-A$ or its normalized form $I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$. Let us define $L = I - W$. The heat kernel rank in Equation (4) is then rewritten as

$$p_t = \sum_{k=0}^{\infty} \frac{(-t)^k}{k!} p_0 L^k = p_0 e^{-tL}. \tag{5}$$

The discrete version of the heat kernel, first introduced in Chung and Yau (1999), is a symmetric version of $e^{-tL}$. Several approximations have been proposed for the heat kernel rank, namely an additive version by taking a finite sum [cf. Equation (4)], or a sum of two infinite products as given by Euler (Chung and Yau, 1999). We rely on the discrete approximation of Yang *et al.* (2007) to compute the vector of ranking scores $p_t$ as in Equation (6)

$$p_t = p_0 \left( I + \frac{-t}{N} L \right)^N, \tag{6}$$

where $N$ is the number of iterations, and $t$, $p_0$ and $L$ are as previously defined. For an error threshold $\varepsilon$, $N$ can be determined such that Equation (7) holds,

$$\left\| p_0 \left( e^{-tL} - \left( I + \frac{-t}{N} L \right)^N \right) \right\| < \varepsilon \tag{7}$$

for any $p_0$ whose entries sum one. Solving Equation (7) is difficult, thus Yang *et al.* (2007) use a heuristic motivated by the following. When $L = I - W$,

$$\left( I + \frac{-t}{N} (I-W) \right)^N = \begin{pmatrix} 1 & * & * \\ 0 & \left(1 + \frac{t(\lambda_2 - 1)}{N}\right)^N & * \\ 0 & 0 & \ddots \end{pmatrix} \tag{8}$$

with $1, \lambda_2, \ldots$ being the eigenvalues of $W$. Note that the eigenvalues of the above matrix are $\left(1 + \frac{t(\lambda-1)}{N}\right)^N - e^{t(\lambda-1)}$. Thus, the heuristic method to determine $N$ considers the difference between positive eigenvalues and a given threshold. For instance, for $t=1$ and $\lambda < 1$, then

$$\left| \left( 1 + \frac{t(\lambda-1)}{N} \right)^N - e^{t(\lambda-1)} \right| < 0.005 \tag{9}$$

for $N \geq 100$. Usually, it is possible to obtain satisfactory ranking results even for small values of $N$, being sufficient to consider $N \simeq 10$.

## 2.2 Customizing TFRank: parameter contribution

TFRank is a flexible integrative prioritization framework allowing for some degree of customization. In particular, parameters may be tuned by experts to guide the method's search according to the problem under study (Fig. 1). We describe the effect of each parameter to the output below.

*2.2.1 Diffusion coefficient* The heat diffusion coefficient parameter, $t$, plays an important role in the heat kernel as it controls the diffusion rate. A large value causes heat to diffuse rapidly, giving preference to regulators located farther away in the network. Conversely, a small value promotes a slow diffusion, thus favoring nearby TFs.

*2.2.2 Preference vector* Biological context is introduced by the preference vector, $p_0$, which defines the initial scores for vertices. Distinct initialization schemes can be used, aiming at differentiating the set of target genes of interest from the remaining genes in the network by attributing them a larger score and therefore greater importance. In general,

$$p_0(u) = \begin{cases} p_{0_T}(u) & \text{if } u \in T, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Common practice is to assume that $p_0$ follows a uniform distribution over the set of target genes $T$, with $p_{0_T}(u) = 1/|T|$, or an equal scoring setting with $p_{0_T}(u) = 1$. Alternatively, genes in $T$ may yield distinct initial scores based on experimental measurements such as expression levels, or based on a correlation with a particular outcome. Input scores may also follow a degree-based distribution with $p_{0_T}(u) = d(u) / \sum_{z \in T} d(z)$.

*2.2.3 Regulation weights* Regulation weights, or edge weights in the network graph, can be used to define the extent of contribution of each TF–target interaction to the diffusion. Typically, there is little information available in this regard, but we believe that incorporating measures of confidence or intensity on these associations is likely to improve the ranking accuracy. The weight $w(u,v)$ of every edge $(u,v)$ can be set as the binding affinity between the TF and the binding site in the promoter region of the target gene, derived by ChIP-chip assays or computational TF–target prediction methods. Note that when using normalized weights, the entries of matrix $W$ in Section 2.1 acquire the form $w(u,v) / \sum_{(u,z) \in E} w(u,z)$. Specifically for yeast using YEASTRACT, $w(u,v)$ can be defined as

$$w(u,v) = w + \sum_{\tau \in \{D,I,U\}} w_\tau \times \delta_\tau(u,v), \tag{11}$$

where predefined constants $w_D$, $w_I$ and $w_U$ denote the weights attributed to regulations pertaining direct ($D$), indirect ($I$) and undefined ($U$) evidence, respectively. Binary function $\delta_\tau(u,v)$ evaluates to 1 when edge $(u,v)$ yields evidence of type $\tau$, or to 0 otherwise, where $\tau \in \{D,I,U\}$. When all edges are equally relevant, $w_\tau = 0$ and Equation (11) is then $w(u,v) = w$.

*2.2.4 Score normalization* In order to attenuate the bias toward genes with a large number of regulatory associations, the final ranking can be refined by degree normalizing the scores: $r_t(u) = p_t(u)/n(u)$. The normalization factor, $n(u)$, can be chosen to be the outdegree, indegree or the sum of both out and in degrees of vertex $u$. When applied to the original network graph, these will mitigate the effect of TFs regulating many genes, genes regulated by many TFs or both, respectively. Weight may be used instead of degree and logarithmic normalizations can also be considered. We note that normalization penalizes only TF first-level regulations.

## 3 RESULTS AND DISCUSSION

In this section, we evaluate the ability of TFRank to provide informative rankings of TFs underlying response to stress in yeast or controlling gene markers of complex diseases in human. We also assess the contribution of each parameter to the output.

## 3.1 Case studies in yeast

We retrieved yeast regulatory data from YEASTRACT (Abdulrehman *et al.*, 2011), yielding 184 regulators and over 48 000 documented TF–target interactions. TFs were mapped to their encoding genes and all TF-encoding and target genes identifiers were mapped to the same *Saccharomyces* genome database (SGD) nomenclature set. The regulatory graph was built by adding an edge per TF–target association from the node representing the TF-encoding gene to the node representing the target gene. Edges were then reversed and their weights calculated using Equation (11). We processed datasets containing genes upregulated in response to two chemical stresses: the first imposed by the antimalarial drug quinine (dos Santos *et al.*, 2009); and the second induced upon exposure to the food preservative acetic acid (Mira *et al.*, 2010b). Both were submitted to TFRank analysis using identical input parameters. The diffusion coefficient $t$ was set to 0.25 to promote a moderate number of layers of transcriptional control. A larger value (closer to 1) would cause TFRank to rapidly reach the TFs at the top of the regulatory network, those exhibiting a considerable amount of known target genes, eventually masking proximal regulators more specifically related to the biological process under study and thus likely to be more relevant to the regulation of the responsive genes. No score normalization was applied, as the selected diffusion coefficient was already expected to attenuate the bias toward TFs with a large number of targets. We decided not to differentiate edge weights and score all regulatory associations equally, irrespective of the underlying experimental evidence. Although the opposite could theoretically generate more reliable results, diminishing the importance of less well-studied interactions could reduce the network to associations supported by TF–promoter binding experiments. More importantly, this would certainly bias the analysis toward prior knowledge by setting aside the least studied TFs and targets. The transcript levels of the TFs or the target genes were not considered relevant in neither of the case studies. This is reasonable, as the activity of a TF is not exclusively based on its transcript levels, and it may also depend on the occurrence of post-translational modifications or on the alteration of its subcellular localization. In fact, experimental evidences gathered so far show that there is a weak correlation between the ratio of induction level of a particular gene in response to a given stress and its contribution for maximal yeast tolerance to such stress. This was reportedly the case for acetic acid, where the genes exerting the strongest protection against the chemical were those less transcriptionally activated (∼2-fold) (Mira *et al.*, 2010b). Results are summarized in Tables 1 and 3 and Figures 2 and 3, while complete rankings are provided in Supplementary Tables S1 and S2. We analyze them throughout the text.

*3.1.1 Yeast response to quinine-induced stress* TFRank reported Adr1, Hap4, Mal33, Gal4, Cat8 and Gis1 as the top six presumable mediators of yeast transcriptional response to quinine (Table 1). This list significantly differs from the top six ranking sorted by the percentage of documented regulatory associations targeting the set of quinine-induced genes (Table 2). It is noteworthy that some TFs highlighted by TFRank exert documented control over a relatively low number of responsive genes. For example, Mal33, ranked in third, is described as a documented regulator of 2.5% of the activated genes (Table 1). However, these TFs regulate other TFs with large percentages of documented targets among the quinine-induced genes, suggesting that indirect control may also present an important contribution (Fig. 2). The group of top six TFs obtained with TFRank sets aside the top six TFs ranked by the percentage of documented targets, each of them being predicted to control the expression of >35% of genes in the dataset (Tables 1 and 2). In particular, Msn2, Sok2, Rpn4 and Yap1 are known to be involved in the regulation of yeast response to several environmental stresses and consequently exert control over a large number of genes. In this context, their classification at the top of the ranking by percentage based on the YEASTRACT network likely reflects a generic, rather than more specific, regulatory role played by these TFs in the yeast transcriptional response to quinine stress. Notably, the single elimination of *MSN2*, *SOK2*, *RPN4* or *YAP1* genes from the yeast cell was found not to increase the susceptibility of yeast

**Table 1.** TFRank-suggested mediators of the yeast genes activated in response to quinine (QN) stress (dos Santos *et al.*, 2009)

| Regulator | DT | PT |
|---|---|---|
| Adr1 | 30.0 | 36.3 (Hms1,Nrg1,Pdr1,Pdr8,Pho4,Pip2,Put3,Tec1,Uga3,Usv1,Yap6) |
| Hap4 | 23.8 | 32.5 (Cat8,Gcn4,Mig2,Ndt80,Phd1,Put3,Yox1) |
| Mal33 | 2.5 | 35.0 (Abf1,Gln3,Hcm1,Met32,Otu1,Swi4,Xbp1) |
| Gal4 | 5.0 | 10.0 (Mga1,Mth1,Rds1,Sfl1,Zap1) |
| Cat8 | 12.5 | 25.0 (Dal81,Hap4,Sip4) |
| Gis1 | 6.3 | 37.5 (Cin5,Mig3,Mss11,Nrg1,Rds1,Tos8,Usv1,Ypr015c) |

DT: documented targets (% of QN-induced genes reportedly targeted by the TF); PT: potential targets (% of QN-induced genes potentially targeted via other TFs). The table shows the six TFs with highest score according to TFRank (complete list in Supplementary Table S1). For each TF, it also shows the percentages of: QN-induced genes reportedly targeted in the network (YEASTRACT), QN-induced genes controlled by other TFs (in brackets) which are targeted by the top TF.

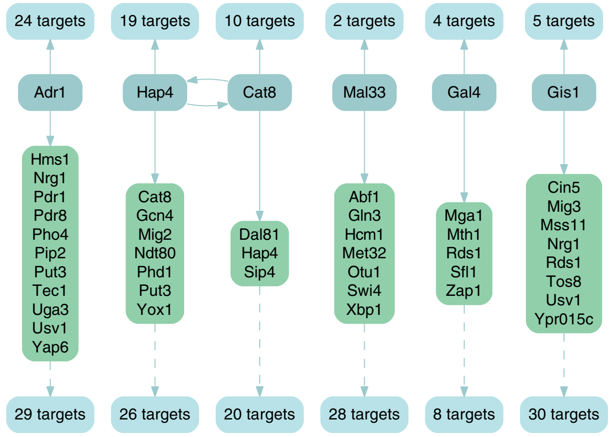

**Fig. 2.** Subnetwork of top six TFs potentially involved in yeast response to quinine-induced (QN) stress, according to TFRank. Light blue boxes show the number of QN-induced genes reportedly regulated by the top TFs (dark blue), or by a layer of other TFs (green) controlled by the top TFs. Figure generated using custom Graphviz (Gansner *et al.*, 2000) code.

**Table 2.** Distribution of the yeast genes activated in response to quinine (QN) (dos Santos *et al.*, 2009) or acetic acid (AA) stress (Mira *et al.*, 2010b) per TF in the top six positions of the list ranked by percentage based on the YEASTRACT network (complete lists in Supplementary Tables S1 and S2).

| Dataset | Top 6 TFs (% of documented targets in the dataset) | | |
|---|---|---|---|
| Quinine | Aft1 (41.3),<br>Ste12 (38.8), | Msn2 (40.0),<br>Rpn4 (36.3), | Sok2 (38.8),<br>Yap1 (35.0) |
| Acetic acid | Haa1 (60.4),<br>Ste12 (45.0), | Sfp1 (53.2),<br>Msn2 (41.4), | Yap1 (50.5),<br>Aft1 (35.1) |

to quinine (dos Santos and Sá-Correia, 2011) and none of them was upregulated in quinine-stressed yeast cells (dos Santos *et al.*, 2009). In contrast, all top six TFs highlighted by TFRank were found upregulated in response to quinine stress (4.1-fold for *ADR1*, 10.6-fold for 3-fold for *HAP4*, *MAL33*, 2.2-fold for *GAL4*, 7.6-fold for *CAT8* and 3.6-fold for *GIS1*) (dos Santos *et al.*, 2009). All these TFs are involved in yeast adaptation to carbon sources alternative to glucose. Adr1 and Cat8 participate in the regulation of genes required for the metabolism of non-fermentable carbon sources (Haurie *et al.*, 2001; Young *et al.*, 2003), whereas Gis1 controls the reprogramming of carbon metabolism during transition to stationary phase (Zhang *et al.*, 2009) and genes involved in the adaptation to nutrient limitation (Pedruzzi *et al.*, 2000). In turn, Hap4, Mal33 and Gal4 control the transcript levels of genes required for respiration (Schüller, 2003), and maltose (Feuermann *et al.*, 1995) and galactose (Bhat and Murthy, 2001) catabolism, respectively. Quinine has been shown to competitively inhibit glucose uptake in yeast cells presumably leading to a state of intracellular glucose limitation, even if glucose is present in the growth medium at saturating concentrations (dos Santos *et al.*, 2009).

*3.1.2 Yeast response to acetic acid-induced stress* TFRank and the method ranking by target percentage of induced genes agreed in the first two TFs for the acetic acid set, Haa1 and Sfp1 (Tables 3 and 2, respectively). Haa1p has been identified as the main regulator of the yeast transcriptional response to this stress (Mira *et al.*, 2010b). The participation of Sfp1 was examined but no detectable effect on cell protection against acetic acid could be attributed to *SFP1* expression (Mira *et al.*, 2010a). TFRank further ranked Msn4, Nrg1 and Fkh2 on top (Table 3). These four TFs target a low percentage of acetic acid-induced genes and therefore appear lower in the list ranked by such criterium (Table 2). Less importance was given by TFRank to Ste12, Msn2, Yap1 and Aft1 regulators, each of them predicted to control the expression of >45% of the acetic acid-responsive genes (Supplementary Table S2 and Table 2, respectively). Microarray experiments revealed upregulation of *MSN4*, *FKH2* and *NRG1* in response to acetic acid stress (Mira *et al.*, 2010b). This is surprising, as the expression was not embedded and thus did not contribute to the TFRank analysis. It suggests that, in some cases, the list of genes could, by itself, hold sufficient information to uncover key TFs. *MSN4*, *FKH2* and *NRG1* genes were also found to confer protection against acetic acid, although at distinct levels (Mira *et al.*, 2010b). Their upregulation in response to acetic acid stress was shown to be dependent on Haa1p (Mira *et al.*, 2010b) (Fig. 3). This tight regulatory scheme seems to be confirmed by the ranking and is an interesting example of the perspective through which TFRank interprets a transcriptional response: as the result of whole-network control rather than isolated action of TFs. Indirect regulation becomes evident in this study, where the top-ranked TFs exert control over other major regulators (Fig. 3). Sfp1 was also highlighted by TFRank and appears to be linked to the remaining players as a target of Fkh2 (Fig. 3). Nevertheless, the elimination of *SFP1* was not found to increase yeast susceptibility to acetic acid stress (Mira *et al.*, 2010a).

**Table 3.** TFRank-suggested mediators of yeast genes activated in response to acetic acid (AA) stress (Mira *et al.*, 2010b)

| Regulator | DT | PT |
|---|---|---|
| Haa1 | 60.4 | 23.4 (Dig1,Fkh2,Msn4,Mth1,Nrg1,Stp3) |
| Sfp1 | 53.2 | 42.3 (Ace2,Adr1,Aft1,Aft2,Arr1,Ash1,Bas1,Cad1,Cin5,Cup9,Gcr1,Gln3,Gzf3,Hac1,Hap1,Hap5,Hms1,Hms2,Hsf1,Ifh1,Ime1,Kar4,Met4,Mga2,Mot3,Nrg2,Otu1,Pdr1,Pdr3,Pho4,Rme1,Rph1,Rpn4,Srd1,Stb5,Stp3,Sut1,Swi4,Swi5,Thi2,Tos8,Uga3,Ume6,Xbp1,Yap1,Yhp1,Yrr1) |
| Msn4 | 38.4 | 44.1 (Adr1,Fhl1,Gat1,Ime1,Imp2,Mss11,Rox1,Rpn4,Yap1) |
| Nrg1 | 21.6 | 55.0 (Cup9,Gat3,Gat4,Hap4,Hms1,Ime1,Imp2,Mga1,Mig3,Mot3,Mth1,Nrg2,Pdr1,Pdr8,Rsf2,Sfl1,Sok2,Usv1,Xbp1,Yap6,Yap7) |
| Fkh2 | 11.7 | 62.2 (Ace2,Cup9,Fkh1,Gat1,Kar4,Rsf2,Sfp1,Sut1,Swi5,Yap6,Yhp1) |

DT: documented targets (% of AA-induced genes reportedly targeted by the TF); PT: potential targets (% of AA-induced genes potentially targeted via other TFs). The table shows the six TFs with highest score according to TFRank (complete list in Supplementary Table S2). For each TF, it shows percentages of: AA-induced genes reportedly targeted or controlled by other TFs (in brackets) targeted by the top TF.

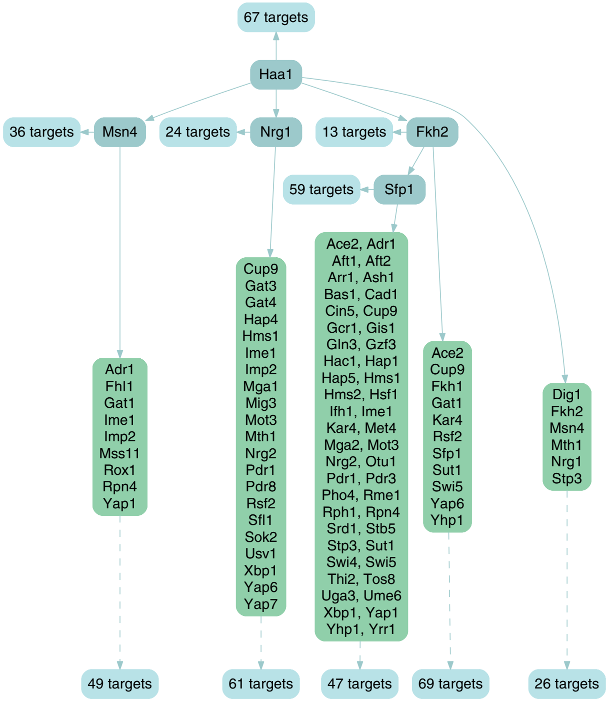
67 targets
Haa1
36 targets
Msn4
24 targets
Nrg1
13 targets
Fkh2
59 targets
Sfp1
Adr1 Fhl1 Gat1 Ime1 Imp2 Mss11 Rox1 Rpn4 Yap1
Cup9 Gat3 Gat4 Hap4 Hms1 Ime1 Imp2 Mga1 Mig3 Mot3 Mth1 Nrg2 Pdr1 Pdr8 Rsf2 Sfl1 Sok2 Usv1 Xbp1 Yap6 Yap7
Ace2, Adr1 Aft1, Aft2 Arr1, Ash1 Bas1, Cad1 Cin5, Cup9 Gcr1, Gis1 Gln3, Gzf3 Hac1, Hap1 Hap5, Hms1 Hms2, Hsf1 Ifh1, Ime1 Kar4, Met4 Mga2, Mot3 Nrg2, Otu1 Pdr1, Pdr3 Pho4, Rme1 Rph1, Rpn4 Srd1, Stb5 Stp3, Sut1 Swi4, Swi5 Thi2, Tos8 Uga3, Ume6 Xbp1, Yap1 Yhp1, Yrr1
Ace2 Cup9 Fkh1 Gat1 Kar4 Rsf2 Sfp1 Sut1 Swi5 Yap6 Yhp1
Dig1 Fkh2 Msn4 Mth1 Nrg1 Stp3
49 targets
61 targets
47 targets
69 targets
26 targets

**Fig. 3.** Subnetwork of TFs presumably involved in yeast response to acetic acid (AA) stress, according to TFRank. Counts of AA genes (light blue) regulated by the top TFs (dark blue) or by TFs (green) controlled by those.

## 3.2 Evaluation of parameter contribution

We assessed the contribution of each TFRank parameter as follows. We observed the impact produced by its variation on the ranks of TFs whose importance would change, in line with the effect described in Section 2 (details in Supplementary Material). For the diffusion coefficient, we used a large or small *t* value, expected to promote distal and proximal TFs, respectively, and two distinct edge weight and seed score schemes. We determined the positions of the 46 TFs with outdegree above the 75th percentile and observed that these (farthest) TFs tended to rank higher when a larger coefficient was used, as expected. We also observed that TFs with reduced outdegree (proximal) ranked higher with the lower $t$, while TFs controlling large regulons (distal) were favored by the larger $t$ value. For the edge weights, we relied on two different schemes expected to benefit indirect and direct evidence, and two distinct diffusion coefficients and two seed score settings. We assessed the ranks of the 16 TFs for which >90% of its TF–target interactions yielded indirect evidence. These explicitly ranked higher with the scheme favoring indirect evidence, confirming our hypothesis. For the preference vector, we compared the rankings obtained using expression values *versus* equal initial scores, with high and low $t$ and two distinct edge weight schemes. We verified the ranks achieved by the 7 and 8 TFs whose encoding genes were upregulated in the quinine and acetic acid sets, respectively. In the quinine set, the group reportedly ranked higher with $t = 0.75$ when using expression relative to equal scores, but it ranked equivalently with $t = 0.25$ using both preference vector settings. The group contained TFs which were also the most proximal and were thus favored by the lower $t$ independently of the initial scoring. In the acetic acid set, most of the eight TFs yielded lower upregulation levels than the remaining genes and therefore could not compete with TFs targeting many genes in the set. We performed an additional test in which we artificially boosted the initial scores of the TFs. These undoubtedly ranked on top, supporting the effect described in Section 2. Finally, for normalization we analyzed the rankings obtained using the regular against normalized score, with two distinct edge weight and two seed score settings. We observed the ranks of the 26 and 13 TFs yielding no targets in the quinine and acetic acid sets, respectively. They achieved higher ranks when sorted based on the normalized score, showing that TFs controlling more genes were effectively penalized.

## 3.3 Preliminary study in human

For human, we collected JASPAR position-specific scoring matrices (PSSMs) and UCSC Human Genome (hg19) sequences made available by Zambelli *et al.* (2009). Each matrix identifier (UniProt) was mapped to its encoding gene (NCBI Entrez) using the official UniProt ID conversion service completed with manual search on databases. RefSeq sequence accession numbers (NCBI GRCh37, February 2009) were converted to Entrez using a file from the NCBI repository. Matrices and sequences with unmapped identifiers were filtered. Sequences were further processed to select the stretches 200 bp upstream and 0 bp downstream the transcription start site. We used the PoSSuM software (Beckstette *et al.*, 2006) to match the PSSMs against the sequences and filtered results below a $P$-value cutoff of $1\times10^{-4}$. Edge weights $w_{uv}$, for edge $(u,v)\in E$ (Section 2), were computed from the raw matching scores output by PoSSuM by rescaling the original score interval of each PSSM to $[0,1]$. To avoid multiple matches, we selected only the match of highest score for every PSSM–sequence pair. This procedure yielded a network containing 50 386 unique TF–target interactions between 18 088 genes, from which 65 acted as TFs. Finally, we selected two gene sets concerning the study of time to distant metastasis in primary breast cancer (van't Veer *et al.*, 2002; Wang *et al.*, 2005). These contained 70 and 76 genes, respectively, whose expression signature was considered predictive of short interval to distant metastasis in lymph node-negative patients. Gene identifiers were first mapped to official HGNC symbols (using a HGNC file or the GeneAnnot service at GeneCards for Affymetrix chip U133A probe set identifiers, respectively) and then converted to Entrez. Duplicates were removed, resulting in 44 and 64 unique genes, respectively. TFRank was applied to the human network with $t = 0.25$, $N = 100$ and no normalization, similarly to the studies in yeast. We took as initial preference vector values: the absolute correlation between the expression signature of each gene and the prognostic category in the case of the van't Veer *et al.* (2002) set, or the absolute standard Cox coefficient in the case of the Wang *et al.* (2005) set.

**Table 4.** Regulators suggested by TFRank and PASTAA to control genes predicting short interval to metastasis in breast cancer: VV (van't Veer *et al.*, 2002) and WA (Wang *et al.*, 2005)

| TFRank-TFs | DT | PT | DTN | PR | PASTAA-TFs | TR |
|---|---|---|---|---|---|---|
| SP1 | 41.0 | 82.1 | 22.1 | 28 | TFAP2A | 12 |
| EWS-FLI1 | 20.5 | 74.4 | 10.8 | 55 | NFKB1 | 19 |
| MZF1 | 12.8 | 94.9 | 4.5 | 16 | NHLH1 | 10 |
| FOXL1 | 10.3 | 79.5 | 4.5 | 20 | CREB1 | 5 |
| CREB1 | 12.8 | 51.3 | 3.7 | 4 | TAL1-TCF3 | 15 |
| IRF1 | 12.8 | 53.8 | 7.2 | 39 | E2F1 | 14 |
| SP1 | 34.0 | 90.0 | 22.1 | 3 | ELK1 | 2 |
| ELK1 | 12.0 | 62.0 | 5.3 | 1 | SRF | 33 |
| NFATC2 | 10.0 | 86.0 | 9.6 | 39 | SP1 | 1 |
| CREB1 | 8.0 | 58.0 | 3.4 | 17 | YY1 | – |
| TFAP2A | 8.0 | 96.0 | 5.2 | 50 | ELK4 | 34 |
| FEV | 14.0 | 0.0 | 6.6 | 64 | JUN-FOS | 46 |

DT: documented targets (% of targets in gene set); PT: potential targets (% of indirect targets of TFs which are targets); DTN: oc. targets in network. We indicate the number of VV and WA genes in the TFRank and PASTAA TF-target sets. For each TF, we include the rank attributed by the competing method (PR-PASTAA; TR-TFRank). The upper part shows the top six ranked TFs for the 44 VV genes (40 in TFRank, 44 in PASTAA). The lower part shows the top six ranked TFs for the 64 WA genes (53 in TFRank, 61 in PASTAA).

### 3.3.1 *Regulation of <5 year-to-metastases biomarkers in human*

Although the gene sets overlapped minimally (one gene), the top six lists output by TFRank shared two TFs, SP1 and CREB1 (Table 4). SP1 was first in both studies. SP1 is known to activate multiple genes involved in tumor cell growth and metastasis (Zhou *et al.*, 2011). EWSR1-FLI1, a fusion protein linked primarily to bone tumor (Ewing's sarcoma), was second for the van't Veer *et al.* (2002) set. Interestingly, a subset of the gene markers from this set has been reported as an indicator of primary breast cancer to osteolytic bone metastasis (Kang *et al.*, 2003). MZF1, ranked in third, is an important transcriptional regulator in hematopoietic development. Unregulated MZF1 expression has been presented as oncogenic and its overexpression seems to induce metastasis through AXL gene expression (Mudduluru *et al.*, 2010). FOXL1, fourth, is a mesenchymal TF with a key role in gastric and colorectal carcinogenesis (Perreault *et al.*, 2005). Notably, it has been found overexpressed in low-grade fibromyxoid sarcoma (Möller *et al.*, 2011), a condition affecting bone, muscular and hematopoietic tissue. CREB1, fifth, has been associated to cancer development and poor prognosis. Recently, Son *et al.* (2010) have reported CREB1 to be overexpressed in advanced breast cancer cells and to influence the expression of several genes involved in metastasis to bone. IRF1, sixth in the list, mediates interferon as well as other cytokine activity, and it has been appointed as a tumor growth inhibitor (Kim *et al.*, 2004). In the Wang *et al.* (2005) dataset, ELK1 ranked second. This factor participates in the chemokine signaling pathway, whose disruption has been implicated in tumor proliferation. ELK1 expression has been found to increase in breast cancer tissue relative to normal cells and to correlate with poor prognosis (Potter *et al.*, 2011). NFATC2, third, is a nuclear factor of activated T cells (NFAT) that integrates the NFAT complex, an inducer of gene transcription during the immune response. NFATC2 is believed to act as a tumor suppressor, but NFAT factors have also been shown to promote breast cancer cell invasive migration by activating COX-2 (Mancini and Toker, 2009). The fourth TF was CREB1, ranked in fifth in the van't Veer *et al.* (2002) set. Upregulation of TFAP2A, ranked in fifth, has been associated with favorable prognosis in breast cancer which conflicts with its role in the positive regulation of the ERBB2 oncogene. Notably, low expression of TFAP2A may cause the loss of tumor suppressor KISS1 and thus increase the risk of metastasis (Mitchell *et al.*, 2006). FEV, sixth, is the second ETS oncogene in the list for the Wang *et al.* (2005) set.

ELK1 and FEV share approximately half of each other targets, including POLQ, whose upregulation presumably correlates with poor prognosis in early breast cancer (Higgins *et al.*, 2010).

*3.3.2 TFRank versus PASTAA* TFRank results were compared with those obtained by PASTAA (Roider *et al.*, 2009) for the same input. PASTAA was selected due to its ability to also consider weighted TF–target links and gene relevance scores. It was previously compared with alternative methods, including PAP (Chang *et al.*, 2006) and an approach based on *z*-score statistics similar to oPOSSUM (Ho Sui *et al.*, 2005). In such study, PASTAA and the *z*-score both yielded superior results, except for one dataset in which only PASTAA seemed to outperform the competing techniques (Roider *et al.*, 2009). In our study, TFRank and PASTAA agreed in placing CREB1 among the top six TFs for the van't Veer *et al.* (2002) set (Table 4). Aside from the TFAP2A and CREB1 factors, and E2F1 whose low expression strongly determines low risk of breast cancer metastasis (Vuaroqueaux *et al.*, 2007), the relation of the remaining TFs in the PASTAA list with breast cancer (and metastasis) was less evident in the literature, where NFKB1, NHLH1 and TAL1-TCF3 were primarily involved in ovary and colorectal (Andersen *et al.*, 2010), brain (De Smaele *et al.*, 2008) and thyroid (Jacques *et al.*, 2009) cancer, respectively. For the Wang *et al.* (2005) set, both methods ranked SP1 and ELK1 in the top 6 TFs. Other TFs promoted by PASTAA included: SRF and YY1, involved in breast cancer cell migration (Lieberthal *et al.*, 2009; Medjkane *et al.*, 2009) ; ELK4 related to prostate cancer (Makkonen *et al.*, 2008); and JUN-FOS, implicated in tumor differentiation and progression (Hein *et al.*, 2009). Overall, TFRank seemed to unravel TFs more specifically related to breast cancer and its metastastatic forms. Particularly considering the van't Veer *et al.* (2002) data, TFRank highlighted a number of TFs presumably involved in breast cancer metastasis to bone, in accordance with the hypothesis that the 70-gene signature predicted bone as an occurrence site (Kang *et al.*, 2003). TFRank also recovered PASTAA key TFs better than vice versa (Table 4).

## 4 CONCLUSION

We presented TFRank, a graph-based framework to identify relevant regulators mediating specific transcriptional responses through whole-network analysis of gene activity. TFRank successfully highlighted regulators potentially involved in the yeast response to quinine- and acetic acid-induced stresses. In a preliminary study in human, TFRank revealed TFs presumably controlling genes related to metastasis in breast cancer. We observed that by assuming a systemic rather than isolated TF action perspective, TFRank captured valuable information about the regulatory mechanisms that would not have been unveiled otherwise. We showed that this strategy provides adequate parameters to effectively control the level of regulatory specificity for the problem under study, fine tune TF–target association scores and incorporate expression levels or alternative measures of relevance into the input gene list. TFRank relies on an additive model to compute the relevance score, which requires additional care when mixing positive and negative input values. A potential workaround is to analyze genes in two distinct groups or to use absolute values. The presented case studies supported TFRank as a promising tool to unveil important TFs, despite the real set of regulations active in the process under study remained undisclosed (all known TF–target interactions were used). Such knowledge could be extracted from experimental ChIP-chip data and used to derive a more accurate network. Envisioned future work includes the integration of TFRank with an algorithm for capturing coherent transcriptional trends across multiple conditions.

*Funding*: Work supported by Fundação para a Ciência e a Tecnologia (FCT): (projects PTDC/AGR-ALI/102608 /2008, PTDC/EBB-EBI/108517/2008, PTDC/EIA-EIA/111239/2009, grants SFRH/BD/36586/2007 to J.P.G., SFRH/BPD/46982/2008 to N.P.M.), and INESC-ID multiannual funding (PIDDAC program).

*Conflict of Interest*: none declared.

## REFERENCES

Abdulrehman,D. *et al.* (2011) YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in Saccharomyces cerevisiae through a web services interface. *Nucleic Acids Res.*, **39**, D136–D140.

Andersen,R. and Lang,K.J. (2006) Communities from seed sets. In *Proceedings of the 15th International Conference on World Wide Web (WWW'06)*. ACM, New York, NY, USA, p. 223.

Andersen,V. *et al.* (2010) Polymorphisms in NFkB, PXR, LXR and risk of colorectal cancer in a prospective study of Danes. *BMC Cancer*, **10**, 484.

Balleza,E. *et al.* (2009) Regulation by transcription factors in bacteria: beyond description. *FEMS Microbiol. Rev.*, **33**, 133–151.

Bar-Joseph,Z. *et al.* (2003) Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.*, **21**, 1337–1342.

Beckstette,M. *et al.* (2006) Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinformatics*, **7**, 389.

Bhat,P.J. and Murthy,T.V.S. (2001) Transcriptional control of the GAL/MEL regulon of yeast saccharomyces cerevisiae: mechanism of galactose-mediated signal transduction. *Mol. Microbiol.*, **40**, 1059–1066.

Booth,L.N. *et al.* (2010) Intercalation of a new tier of transcription regulation into an ancient circuit. *Nature*, **468**, 959–963.

Brin,S. and Page,L. (1998) The anatomy of a large-scale hypertextual Web search engine. *Comput. Netw ISDN*, **30**, 107–117.

Chang,L.-W. *et al.* (2006) A systematic model to predict transcriptional regulatory mechanisms based on overrepresentation of transcription factor binding profiles. *Genome Res*, **16**, 405–413.

Chung,F. (2007) The heat kernel as the pagerank of a graph. *Proc. Natl Acad. Sci. USA*, **104**, 19735–19740.

Chung,F. and Yau,S. (1999) Coverings, heat kernels and spanning trees. *Electr. J. Comb.*, **6**, R12.

De Smaele,E. *et al.* (2008) An integrated approach identifies Nhlh1 and Insm1 as sonic hedgehog-regulated genes in developing cerebellum and medulloblastoma. *Neoplasia*, **10**, 89–98.

dos Santos,S.C. *et al.* (2009) Transcriptomic profiling of the Saccharomyces cerevisiae response to quinine reveals a glucose limitation response attributable to drug-induced inhibition of glucose uptake. *Antimicrob. Agents Chemother.*, **53**, 5213–5223.

dos Santos,S.C. and Sá-Correia,I. (2011) A genome-wide screen identifies yeast genes required for protection against or enhanced cytotoxicity of the antimalarial drug quinine. *Mol. Genet. Genomics*, in press [Epub ahead of print, doi: 10.1007/s00438-011-0649-5].

Essaghir,A. *et al.* (2010) Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data. *Nucleic Acids Res.*, **38**, e120.

Feuermann,M. *et al.* (1995) Sequence of a 9.8 kb segment of yeast chromosome ii including the three genes of the mal3 locus and three unidentified open reading frames. *Yeast*, **11**, 667–672.

Gansner,E.R. *et al.* (2000) An open graph visualization system and its applications to software engineering. *Softwr. Pract. Exp.*, **30**, 1203–1233.

Haurie,V. *et al.* (2001) The transcriptional activator Cat8p provides a major contribution to the reprogramming of carbon metabolism during the diauxic shift in Saccharomyces cerevisiae. *J. Biol. Chem.*, **276**, 76–85.

Heer,J. *et al.* (2005) Prefuse: a toolkit for interactive information visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'05).* ACM, New York, NY, USA, pp. 421–430.

Hein,S. *et al.* (2009) Expression of Jun and Fos proteins in ovarian tumors of different malignant potential and in ovarian cancer cell lines. *Oncol. Rep.*, **22**, 177–183.

Hestand,M.S. *et al.* (2008) CORE_TF: a user-friendly interface to identify evolutionary conserved transcription factor binding sites in sets of co-regulated genes. *BMC Bioinformatics*, **9**, 495.

Higgins,G. *et al.* (2010) Overexpression of POLQ confers a poor prognosis in early breast cancer patients. *Oncotarget*, **1**, 175.

Ho Sui,S. *et al.* (2005) oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Res.*, **33**, 3154–3164.

Jacques,C. *et al.* (2009) Death-associated protein 3 is overexpressed in human thyroid oncocytic tumours. *Br. J. Cancer*, **101**, 132–138.

Kang,Y. *et al.* (2003) A multigenic program mediating breast cancer metastasis to bone. *Cancer Cell*, **3**, 537–549.

Kim,P. *et al.* (2004) IRF-1 expression induces apoptosis and inhibits tumor growth in mouse mammary cancer cells in vitro and in vivo. *Oncogene*, **23**, 1125–1135.

Kim,S.-Y. and Kim,Y. (2006) Genome-wide prediction of transcriptional regulatory elements of human promoters using gene expression and promoter analysis data. *BMC Bioinformatics*, **7**, 330.

Lemon,B. and Tijian,R. (2000) Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev.*, **14**, 2551–2569.

Lieberthal,J.G. *et al.* (2009) The role of YY1 in reduced HP1alpha gene expression in invasive human breast cancer cells. *Breast Cancer Res.*, **11**, R42.

Makkonen,H. *et al.* (2008) Identification of ETS-like transcription factor 4 as a novel androgen receptor target in prostate cancer cells. *Oncogene*, **27**, 4865–4876.

Mancini,M. and Toker,A. (2009) NFAT proteins: emerging roles in cancer progression. *Nat. Rev. Cancer*, **9**, 810–820.

Margolin,A.A. and Califano,A. (2007) Theory and limitations of genetic network inference from microarray data. *Ann. NY Acad. Sci.*, **1115**, 51–72.

Marstrand,T.T. *et al.* (2008) Asap: a framework for over-representation statistics for transcription factor binding sites. *PLoS One*, **3**, e1623.

Martínez-Antonio,A. (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr. Opin. Microbiol.*, **6**, 482–489.

Medjkane,S. *et al.* (2009) Myocardin-related transcription factors and SRF are required for cytoskeletal dynamics and experimental metastasis. *Nat. Cell Biol.*, **11**, 257–268.

Mira,N.P. *et al.* (2010a) Genome-wide identification of Saccharomyces cerevisiae genes required for tolerance to acetic acid. *Microb. Cell Fact.*, **9**, 79.

Mira,N.P. *et al.* (2010b) Genomic expression program involving the Haa1p-regulon in Saccharomyces cerevisiae response to acetic acid. *OMICS*, **14**, 587–601.

Mitchell,D.C. *et al.* (2006) Regulation of KiSS-1 metastasis suppressor gene expression in breast cancer cells by direct interaction of transcription factors activator protein-2alpha and specificity protein-1. *J. Biol. Chem.*, **281**, 51–58.

Möller,E. *et al.* (2011) FUS-CREB3L2/L1-positive sarcomas show a specific gene expression profile with upregulation of CD24 and FOXL1. *Clin. Cancer Res.*, **17**, 2646–2656.

Mudduluru,G. *et al.* (2010) Myeloid zinc finger 1 induces migration, invasion, and in vivo metastasis through Axl gene expression in solid cancer. *Mol. Cancer Res.*, **8**, 159–169.

Pedruzzi,I. *et al.* (2000) Saccharomyces cerevisiae Ras/cAMP pathway controls post-diauxic shift element-dependent transcription through the zinc finger protein Gis1. *EMBO J.*, **19**, 2569–2579.

Perreault,N. *et al.* (2005) Foxl1 is a mesenchymal Modifier of Min in carcinogenesis of stomach and colon. *Genes Dev.*, **19**, 311–315.

Potter,S.M. *et al.* (2011) Influence of stromal-epithelial interactions on breast cancer in vitro and in vivo. *Breast Cancer Res. Treat.* [Epub ahead of print, doi:10.1007/s10549-011-1410-9].

Pournara,I. and Wernisch,L. (2007) Factor analysis for gene regulatory networks and transcription factor activity profiles. *BMC Bioinformatics*, **8**, 61.

Reverter,A. *et al.* (2010) Regulatory impact factors: unraveling the transcriptional regulation of complex traits from expression data. *Bioinformatics*, **26**, 896–904.

Roider,H.G. *et al.* (2009) PASTAA: identifying transcription factors associated with sets of co-regulated genes. *Bioinformatics*, **25**, 435–442.

Sandelin,A. *et al.* (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.

Schüller,H.-J. (2003) Transcriptional control of nonfermentative metabolism in the yeast Saccharomyces cerevisiae. *Curr. Genet.*, **43**, 139–160.

Son,J. *et al.* (2010) cAMP-response-element-binding protein positively regulates breast cancer metastasis and subsequent bone destruction. *Biochem. Biophys. Res. Commun.*, **398**, 309–314.

Teixeira,M.C. *et al.* (2006) The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucleic Acids Res.*, **34**, D446–D451.

Teixeira,M.C. *et al.* (2010) Refining current knowledge on the yeast FLR1 regulatory network by combined experimental and computational approaches. *Mol. Biosyst.*, **6**, 2471–2481.

Vallet-Gely,I. *et al.* (2007) Local and global regulators linking anaerobiosis to cupA fimbrial gene expression in Pseudomonas aeruginosa. *J. Bacteriol.*, **189**, 8667–8676.

van't Veer,L.J. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Veerla,S. *et al.* (2010) Genome-wide transcription factor binding site/promoter databases for the analysis of gene sets and co-occurrence of transcription factor binding motifs. *BMC Genomics*, **11**, 145.

Vuaroqueaux,V. *et al.* (2007) Low E2F1 transcript levels are a strong determinant of favorable breast cancer outcome. *Breast Cancer Res.*, **9**, R33.

Wang,Y. *et al.* (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, **365**, 671–679.

Wingender,E. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.

Yang,H. *et al.* (2007) DiffusionRank: a possible penicillin for web spamming. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07).* ACM, New York, NY, USA, p. 431.

Young,E.T. *et al.* (2003) Multiple pathways are co-regulated by the protein kinase Snf1 and the transcription factors Adr1 and Cat8. *J. Biol. Chem.*, **278**, 26146–26158.

Zambelli,F. *et al.* (2009) Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. *Nucleic Acids Res.*, **37**, W247–W252.

Zhang,Y. *et al.* (2008) Network motif-based identification of TF-target relationships by integrating multi-source biological data. *BMC Bioinformatics*, **9**, 203.

Zhang,N. *et al.* (2009) Gis1 is required for transcriptional reprogramming of carbon metabolism and the stress response during transition into stationary phase in yeast. *Microbiology*, **155**, 1690–1698.

Zhou,W. *et al.* (2011) Leptin pro-angiogenic signature in breast cancer is linked to IL-1 signalling. *Br. J. Cancer*, **104**, 128–137.