BIOINFORMATICS

Vol. 30 ECCB 2014, pages i415–i421
doi:10.1093/bioinformatics/btu448

Identifying transcription factor complexes and their roles

Thorsten Will and Volkhard Helms*

Center for Bioinformatics, Campus Building E2.1, Saarland University, D-66123 Saarbrücken, Germany

ABSTRACT

Motivation: Eukaryotic gene expression is controlled through molecular logic circuits that combine regulatory signals of many different factors. In particular, complexation of transcription factors (TFs) and other regulatory proteins is a prevailing and highly conserved mechanism of signal integration within critical regulatory pathways and enables us to infer controlled genes as well as the exerted regulatory mechanism. Common approaches for protein complex prediction that only use protein interaction networks, however, are designed to detect self-contained functional complexes and have difficulties to reveal dynamic combinatorial assemblies of physically interacting proteins.

Results: We developed the novel algorithm DACO that combines protein–protein interaction networks and domain–domain interaction networks with the cluster-quality metric cohesiveness. The metric is locally maximized on the holistic level of protein interactions, and connectivity constraints on the domain level are used to account for the exclusive and thus inherently combinatorial nature of the interactions within such assemblies. When applied to predicting TF complexes in the yeast *Saccharomyces cerevisiae*, the proposed approach outperformed popular complex prediction methods by far. Furthermore, we were able to assign many of the predictions to target genes, as well as to a potential regulatory effect in agreement with literature evidence.

Availability and implementation: A prototype implementation is freely available at https://sourceforge.net/projects/dacoalgorithm/.

Contact: volkhard.helms@bioinformatik.uni-saarland.de

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

1 INTRODUCTION

The complexity of higher eukaryotes cannot be explained without considering the combinatorial regulation of transcription (Levine and Tjian, 2003). Recent experimental work showed that the information encoded by many regulatory control elements, namely specific DNA-binding transcription factors (TFs) and their cofactors, epigenetic marks, microRNAs and other input signals, is integrated to orchestrate the expression patterns of genes in a context-dependent manner in time and space (Spitz and Furlong, 2012).

On a conceptual level, this complex signal processing task is accomplished by an interplay of several control layers. Cis-regulatory modules, functional sequence segments whose accessibility is dependent on the packing state of the chromatin, enable basic logic operations on the input that are essentially implemented by protein–protein interactions between DNA-binding proteins with suitably positioned binding sites (Buchler *et al.*, 2003; Istrail and Davidson, 2005). When bound to the DNA, these proteins often mediate further physical interactions that provide an additional scaffold besides the regulatory sequence. This important aspect expands the current understanding of cis-regulatory signal integration beyond the linear sequence-based code (Junion *et al.*, 2012; Siggers and Gordan, 2014). The scaffold is used to recruit additional TFs or cofactors that entail regulatory capabilities such as epigenetic alterations or mediate interactions with other TFs on selected sites where binding motifs obey certain distance constraints (Diez *et al.*, 2014; Göke *et al.*, 2011).

TFs that participate in such multi-protein complexes to produce decisive output signals are referred to as being 'cooperative' (Aguilar and Oliva, 2008; Spitz and Furlong, 2012). Cooperative TFs are found in shorter distance and are more clustered within the protein interaction network than expected by chance (Aguilar and Oliva, 2008; Manke *et al.*, 2003). Although they influence similar groups of target genes, what supports their role as important regulatory drivers, they neither seem to share similar regulatory inputs nor regulate each other (Aguilar and Oliva, 2008). Recent research showed that cooperative binding events are evolutionary much stronger conserved (Göke *et al.*, 2011; He *et al.*, 2011; Kazemian *et al.*, 2013) and show a greater impact on expression compared with individual binding events (Hemberg and Kreiman, 2011). Furthermore, they turned out to be driving regulators in essential eukaryotic control processes such as the cell cycle in yeast (Simon *et al.*, 2001), body part formation in *Drosophila* (He *et al.*, 2011; Kazemian *et al.*, 2013) or mammalian cell fate determination (Göke *et al.*, 2011; Hochedlinger and Plath, 2009; Wilson *et al.*, 2010). The understanding of this combinatorial interplay requires a paradigm shift from single key players to the coupled activity of many regulatory control elements.

At a first glance, the prediction of protein complexes is a well-established field. Data for physical interactions between proteins in the form of protein–protein interaction networks (PPIN) is abundant and so are suitable clustering approaches that find dense areas in networks. A very successful recent method is ClusterONE (Nepusz *et al.*, 2012), which locally optimizes the cluster quality measure cohesiveness. This is done iteratively by stepwise addition/removal of the locally most valuable incident/boundary proteins. The cohesiveness f of a set of proteins V in a network is generally defined as

$$f(V) = \frac{w^{in}(V)}{w^{in}(V) + w^{bound}(V)}$$

where $w^{in}(V)$ denotes the total weight of edges between members of V (internal edges), and $w^{bound}(V)$ denotes the total weight of edges that connect V with the rest of the network (boundary edges). Figure 1 illustrates these definitions with an example and additionally introduces the notion of incident and boundary proteins. Cohesiveness assesses the structural properties of subnetworks that we aim for in this work: they should be densely connected but at the same time well separated from the outside.

*To whom correspondence should be addressed.

An important feature of ClusterONE and other successful complex prediction approaches is to account for overlaps between complexes. Many protein complexes are organized in a modular and combinatorial fashion so that proteins may take part in several distinct complexes (Han *et al*., 2004; Jung *et al*., 2010; Kim *et al*., 2006). PPINs provide a static compilation of assumed pairwise interactions, whereas real biological interactions are highly dynamic and intrinsically controlled by protein expression and spatial constraints. To enable the formation of a physically interacting protein complex, all involved proteins must be expressed at the same time (Han *et al*., 2004; Jansen *et al*., 2002) and must be capable of forming a stable binding topology devoid of any binding site competition (Keskin and Nussinov, 2007; Kim *et al*., 2006). The integration of additional data should thus allow us to derive a clearer picture of the combinatorial complexity of protein complexes.

Although the combinatorial manifold of individual proteins was studied before (Tuncbag *et al*., 2009; Tyagi *et al*., 2009), few computational methods so far account for those effects. A first step in this direction was taken by Jung *et al*. (2010) who annotated mutually exclusive interactions in a yeast protein network using structural data, where available, to determine all possible conflict-free subnetworks. Owing to the exponential complexity, the construction was limited to previously determined dense areas of the network that were then used to predict complexes. The evaluation showed that the effort can lead to a considerable refinement of a given pre-clustering by the consequent exclusion of superfluous proteins from complex candidates. To our knowledge, this is the only available method that accounts for the combinatorial possibilities due to binding interface limitations.

Due to the current sparsity of available structural data on the 3D conformations of protein–protein complexes, a model based on domain–domain interactions (DDIs) as introduced by Ozawa *et al*. (2010) appeared to us as a worthwhile alternative. Given a dense protein interaction subnetwork from a generic complex prediction approach, one decomposes the member proteins into their domains and bases the connectivity on interactions between the individual domains, the DDIs. This transformation to a domain-level network can then be used to filter out false-positive predictions. If every domain is restricted to participate in only one interaction, the model implies a certain binding interface constraint and thus reveals which proteins can be connected at the same time. This is important because true protein complexes must admit a topology of pairwise binding at the level of interfaces, and these interfaces are often exclusive (Aloy and Russell, 2006; Kim *et al*., 2006). The model assumes that the correct topology has the maximal number of simultaneous interactions. Ma *et al*. (2012) enhanced the initial model by introducing a sequence-based domain prediction step and artificial domains that help to maintain connectivity due to incomplete annotation. Furthermore, they suggested a polynomial time solution. Deploying the DDI model as a filtering step to existing clustering approaches was shown to increase the precision of predictions. However, the methods by Ozawa and Ma only provide a fixed solution for each complex candidate. If one is interested in the combinatorial manifold of potential solutions, one would have to enumerate all possible exclusive choices in advance. This is the gap we want to fill with our algorithm.

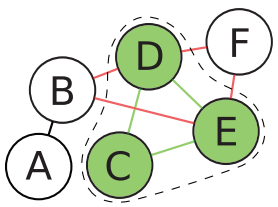

Fig. 1. For convenience, all edges in the PPIN have unit weight and the corresponding weight annotation is omitted. The current members of the cohesive subset $V = \{C, D, E\}$ and their internal edges are colored green, boundary edges are marked red. Boundary edges can be thought to span a boundary (shown as a dashed line) that separates the current dense subset V from the remaining network. This border defines the set of incident proteins $V_{inc} = \{B, F\}$, external vertices adjacent to the boundary, and boundary proteins $V_{bound} = \{D, E\}$, internal vertices at the boundary. For the given example $w^{in}(V) = 3$, $w^{bound}(V) = 4$ and $f(V) = \frac{3}{7}$

2 MATERIALS AND METHODS

We developed a novel combinatorial approach that combines the local optimization of cohesiveness based on weighted protein interaction data and the competition of binding interfaces by adapted integration of domain interaction data. We consequently termed the method 'domain-aware cohesiveness optimization' or 'DACO'. The integration of different levels of granularity coupled with the set of rules introduced by the DDI model is used to enforce a loop invariant within the iterative algorithm that ensures a conflict-free domain interaction topology within the currently grown dense subset during the execution. Hence, a selection of densely connected proteins is always accompanied with a valid spanning tree on the domain level.

A prototype of the approach was implemented in Python (Fig. 2). The only data sources needed as input are a probability-weighted PPIN in the SIF format (simple interaction format, interaction partners and weight are supplied linewise) with nodes named by their UniProt IDs (The UniProt Consortium, 2014), a list of proteins that are used to seed the growth, a threshold to generate the seed pairs, and an upper bound for

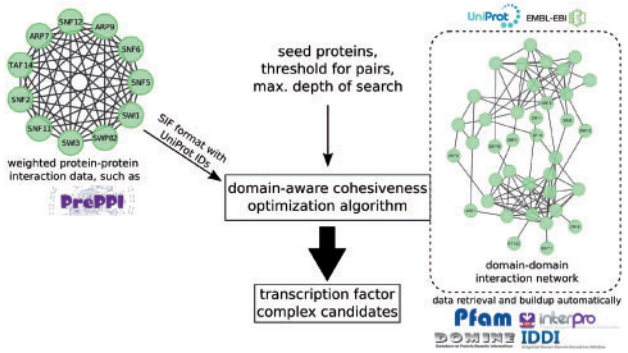

Fig. 2. Shown is the workflow of DACO with its necessary input data, automatically retrieved information and computed output. The network examples show the corresponding subnetworks of the SWI/SNF complex [as annotated in CYC2008 (Pu *et al*., 2009)] built within the documented framework and visualized using Cytoscape (Saito *et al*., 2012). The additional domain information (right) enables a different view on the connectivity of the complex compared with the perfect clique connected by high-probability edges in the PPIN (left)

the depth of search to keep the combinatorial explosion on a local level. Given those requirements, the tool automatically retrieves all the data that it needs to build the DDIN and executes the proposed DACO. The program queries Pfam (Punta *et al.*, 2012) and InterPro (Hunter *et al.*, 2012) for the domain annotation and IDDI (Kim *et al.*, 2012) and DOMINE (Yellaboina *et al.*, 2011) as the data sources for domain interactions. Furthermore, UniProt annotations and the 'dbfetch' Web service at EMBL-EBI (Lopez *et al.*, 2003) are used for name conversion tasks. The implementation is completely independent of the organism and can directly be applied to any other desired organism if the required input data are provided. Details are given in the Supplementary Materials (Section 1.3).

2.1 Domain-aware cohesiveness optimization

Figure 3 illustrates the definitions for incident and boundary proteins on the domain level. Given a certain execution state of the algorithm, defined as the set of currently selected proteins in the assumed cluster V and the active domain interactions between them, the first step is to determine incident and boundary proteins. Connectivity issues are considered according to the domain occupancy of the DDIN using the constraints of the DDI model. With the introduction of this abstraction, binding site conflicts as defined by the model are precluded right from the start to retain the self-imposed loop invariant.

In the next step, we check for all incident proteins whether the cohesiveness can be increased when adding them to V. The same test is applied for removing all boundary proteins. This computation is solely performed on the level of the weighted protein interactions and is independent of the current domain occupancy in the sense that all interactions are taken into account. Although individual edge weights may be unreliable, deviations will average out when many weighted interactions are combined (Nepusz *et al.*, 2012). Among all possible modifications V' we select the one that maximizes the cohesiveness. Every iteration can have three outcomes: the algorithm could terminate because no further increase is possible, the removal of a protein could yield the largest gain or the addition of a protein may be the optimal choice. In the first case, the current complex candidate set V is returned and the algorithm terminates, as the cohesiveness is locally optimal. The removal case changes the current state. The boundary protein P is removed from V leading to $V' = V \backslash \{P\}$. Additionally, the domains occupied by the distinct spanning edge that connected P to the remaining cluster V' are made available again. The definition of boundary proteins ensures the preservation of connectivity within domains of V'. The next iteration is then computed using this modified state.

Adding a protein P is more difficult. Certainly, V is modified to $V' = V \cup \{P\}$ and P needs to establish a domain interaction with a protein within V. But often several interactions are able to link P to V on the domain level. This choice can be crucial for further expansions because it may lead to differing occupancies on the domain level and therefore changes the moves that will be possible in later steps. To take this into account, the algorithm branches and evaluates the outcome of every possible spanning edge on the domain level. The DDIN does not provide a qualitative rating for its individual edges, and all choices lead to the same increase in cohesiveness. However, a reasonable consideration is to exploit the weight annotation of the protein interaction network. Associating each domain interaction with the probability of the corresponding edge between the connected proteins in the PPIN establishes an overall probability that can be used as an early pruning criterion. This correspondence will likely represent an upper limit achievable by any of the domain interactions between the proteins (or even a combination of several) and thus provides a conservative estimate. When assuming that spanning edges in the DDIN are independent of each other, the probability to observe an underlying spanning tree on the domain level is the product of the probabilities of all its spanning edges. If, at a particular iteration of the algorithm, a resulting cohesiveness-maximizing candidate would yield a value <0.5, the current cluster is returned instead. In the case of an addition, all alternatives below the threshold are withdrawn early and the current cluster is only returned if no alternative choice shows a total probability within the boundary. Further information and pseudocode is provided in the Supplementary Materials (Section 1.1/1.2).

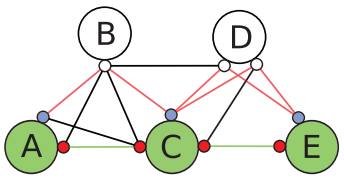

Fig. 3. The green nodes in this DDIN are the proteins currently included in $V = \{A, C, E\}$ and the green edges show how the current dense cluster is connected on the domain level. Incident nodes $V_{inc} = \{B, D\}$ are those nodes that can be connected to V by a new domain-interaction edge (colored in red) to an unused domain of an internal protein (blue domains). Boundary nodes $V_{bound} = \{A, E\}$ are proteins in V with only one used domain. A later removal of an internal node with two or more occupied domains would inevitably lead to a breakdown of the spanning tree on the domain level and thus introduce inconsistencies during execution

Local cohesiveness optimization in ClusterONE starts its greedy growth process from single proteins in the network. Like every greedy algorithm, it is prone to local extrema. In our case, this is desirable because we want to grasp the local combinatorial manifold of complexes around TFs and, owing to the implemented branching principle, the algorithm is able to diversify its intermediate states with a variety of justified directionalities induced by the previous history of domain choices. However, each time the complex candidate is enlarged, only one protein is selected for expansion to keep the runtime convenient. This is especially critical and error-prone during the very first expansion step when only one internal edge and the boundary edges of the two proteins taken from inherently noisy network data (von Mering *et al.*, 2002) are taken into account (Supplementary Fig. S2). Considering this, we decided to start the growth process from pre-built pairs of proteins to spice up the optimization. This overcomes any unfounded directional bias early on in the first expansion step and paves the way for the reasoned bias owing to domain constraints. Furthermore, the pairings should be determined on the basis of the probability of a protein interaction between the two proteins up to a certain likelihood if such data are available. While cohesiveness is undoubtedly a powerful measure of cluster quality if several proteins are involved, its included notion of seclusiveness only has a limited validity for pairs. It is especially misleading for proteins that are combinatorically active and thus potentially exhibit higher boundary weights within the cohesiveness calculations, putting them at a disadvantage compared with less promiscuous nodes in the network.

2.2 Setup for yeast computations

We compared our DACO approach to the most recent versions of the popular complex prediction tools MCODE (Bader and Hogue, 2003), MCL (Enright *et al.*, 2002) and ClusterONE (Nepusz *et al.*, 2012) using common benchmarks for protein complex prediction in yeast and the weighted high-quality yeast protein interaction network PrePPI (Zhang *et al.*, 2013).

Figure 2 illustrates the workflow of the DACO algorithm. Given the PrePPI network and 148 TFs as annotated in the Yeast Promotor Atlas (YPA) (Chang *et al.*, 2011), the complete DDIN was built and all seed pairs were selected that exceed a probability of $t = 0.75$ according to PrePPI (or at least two partners if no interaction was within the threshold). Two TFs annotated in the YPA, MATA1 and MAL63 are not

represented in the PrePPI-PPIN and were therefore omitted. The thus generated start seed comprised 1526 distinct protein pairs that form 1898 start states considering the choices of the necessary domain interaction between each pair. The runs were then executed with a depth threshold of $d = 10$, i.e. the DACO algorithm considered complexes containing up to 10 proteins. The parameter choice of t and d is discussed in the Supplementary Materials (Section 2.1).

MCODE (MCD) was used as a Cytoscape3 plug-in (Saito *et al.*, 2012) in version 1.4b2. For MCL, the stand-alone binary in version 12-068 was used. ClusterONE results were generated with the stand-alone implementation in version 1.0. ClusterONE allows the user to manually influence seed node or sets of seed nodes. In the unsupervised mode (termed Cl1 here), the implementation takes care of the initialization. Additionally, two different case-specific seeded variants were set manually. In the first case, all YPA-annotated TFs were set as initial start proteins (Cl1s). In the second case, all pairs used to initialize DACO were also taken to commence ClusterONE (Cl1ps) to assess the benefit of an induced combinatorial flavor. To restrict the results to TF complexes, the output of MCODE, MCL and ClusterONE was filtered for predicted candidates that contained at least one TF annotated in the YPA. Because the compared general approaches allow tuning of individual parameters, optimization of the most influential settings was conducted to obtain the most competitive individual and overall parameter sets for every single score. As the best overall parameter set, we used the one with the largest sum of averaged reference and biological scores (Supplementary Materials Section 2.2).

3 RESULTS AND DISCUSSION

Table 1 lists the quantities of predicted TF-containing complexes and distinct combinations of TFs that are predicted by the individual methods. Our new DACO algorithm suggested 8–85 times more TF complex candidates than any of the established methods. As expected, the ClusterONE run initialized with the pre-built pairs returned the second most TF complexes and variants. MCODE detected the smallest number of complexes. Also, the MCODE output size deviated the most between overall best and individual best parameter set. The differences between the start variants for ClusterONE (that also locally optimizes the cohesiveness but does not use the domain model) were surprising: when initialized with the individual TFs as seeds, fewer complexes were predicted than when starting the growth from hub proteins in the PPIN, irrespective of the type of the hub proteins. Likewise worth mentioning is the comparably small number of candidates suggested when starting from the 1526 pre-built pairs in comparison with the domain-aware approach. Although initiated with seeds that should favor the combinatorial manifold, the number of predicted TF variants was below the number of annotated TFs (148). This suggests that an approach that only uses protein interaction data is not sufficient to grasp the biological borders of such highly modular subnetworks that fulfill binding-interface constraints.

Table 1. Predicted TF complexes and TF complex variants (how many distinct combinations of TFs are involved in complexes) that were obtained by various approaches

	DACO	Cl1ps	Cl1s	Cl1	MCD	MCL
TF complexes	1375	175/176	61/63	106/106	16/38	75/79
TF variants	412	134/138	59/61	80/80	16/38	75/79

Note: Where parameter tuning was conducted, the left value represents the result of the parameter set that gave the overall best balanced performance for all criteria and the right value is the highest value achieved by any parameter set for this criterion.

3.1 Common complex prediction benchmarks

We computed established measures for the quality of complex predictions based on the agreement with reference datasets. Precision, recall and F-score (Li *et al.*, 2010) were calculated on the basis of the overlap score (Bader and Hogue, 2003). Furthermore, we computed the geometric accuracy (Brohée and van Helden, 2006) and the maximum matching ratio (Nepusz *et al.*, 2012). All measures were independently evaluated on three different protein complex reference datasets for yeast, namely CYC2008 (Pu *et al.*, 2009), MIPS (Mewes *et al.*, 2006) and the SGD (Cherry *et al.*, 2012). With the exception of the precision, all quality metrics were calculated in compliance with the corresponding reference complex sets filtered to the subset of known complexes that involve at least one annotated TF. The precision calculation was allowed to match a candidate to a reference complex that does not include a TF. This was done to facilitate matches that potentially recruited larger complexes of regulatory function in yeast, such as SWI/SNF or RSC. In this context, a method that predicts a candidate with high overlap to such a complex and contains at least one TF should not be placed at a disadvantage. The threshold of tests based on the overlap score was set to $\omega(A, B) > 0.25$.

Figure 4 shows the results of all benchmarks for the overall best parameter set of the compared methods. The overall qualitative picture is neither affected if the best parameter set per measure is considered (Supplementary Table S9) nor by the

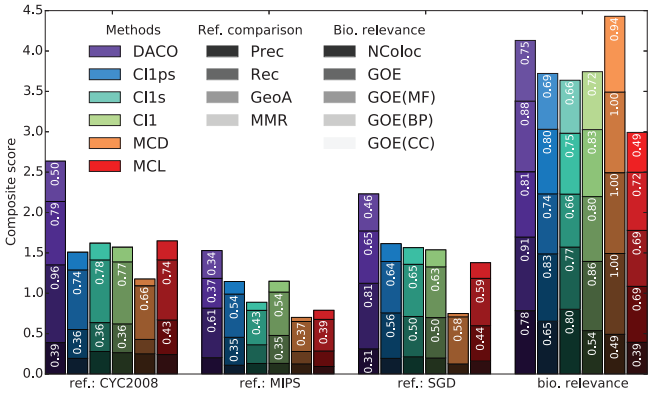

Fig. 4. Graphical summary of the performance of complex prediction methods on reference complex data from CYC2008, MIPS and the SGD as well as measures of biological relevance (Supplementary Table S9). For the parameter-adjusted methods, each evaluation shows the result of the overall best parameter set. The composite score is taken as the sum of the individual quality measures given in the bars. Abbreviated performance measures: Prec: Precision, Rec: Recall, GeoA: geometric accuracy, MMR: maximum matching ratio, NColoc: nucleus colocalization and GOE(x): overrepresentation score based on GO term enrichment (all terms, MF: molecular function, BP: biological process, CC: cellular component)

relaxed definition of the precision (Supplementary Table S11). Except for the geometric accuracy on the MIPS dataset, our novel DACO approach designed for the combinatorial task of predicting TF complexes outperformed the established general algorithms in the agreement with reference complexes. Owing to its local enumeration and thus smaller predicted complexes, DACO is put at a technical disadvantage toward the relatively large TF complexes in MIPS though (Supplementary Table S12). Surprisingly, ClusterONE when started from the curated pairs was on average not superior to the completely unsupervised version. Even the recall was only slightly better for one dataset, whereas the precision was strictly lower. Only in the maximum matching ratio, owing to its missing penalization of non-matching predictions, it performed slightly better. This shows that growing from pairs does not benefit the local cohesiveness optimization as implemented by ClusterONE. This can again be attributed to the missing information of physical constraints in protein interaction networks leading to the fusion of highly modular dense subnetworks (Supplementary Fig. S3). The additional incorporation of the domain topology in the fashion of Ma *et al.* (2012) can be beneficial but leads to an even smaller ensemble of solutions (Supplementary Tables S10 and 11).

Next, we tested the biological plausibility of the results on the basis of colocalization and functional homogeneity within complexes. In the special case of TF complexes, one should expect an *in vivo* localization to the nucleus for all proteins within the same complex. The nucleus colocalization score is defined as the average fraction of proteins per complex encountered in the nucleus weighted by the size of the complex (Friedel *et al.*, 2009). Localization data were taken from Huh *et al.* (2003). Homogeneity was tested as the fraction of complex candidates with at least one enriched Gene Ontology annotation (Ashburner *et al.*, 2000) at significance level $P = 0.05$ (Bonferroni corrected) (Zhang *et al.*, 2008).

Our novel method also performed well in the assessments of biological relevance. While MCODE delivered the best-scoring candidates with respect to enrichment, on average less than half of the proteins among the TF complexes predicted by MCODE are found in the nucleus. MCL was not able to achieve >43% nucleus colocalization with any setting. For the particular task of TF complex prediction, the nucleus colocalization score should be clearly seen as the most important one among the measures of biological relevance. Hence, the DACO approach also clearly succeeds in this aspect.

3.2 Regulatory role of TF complexes

In the next step, we determined the target genes of all involved TFs from the binding data provided by the YPA. In total, 79% of the 412 distinct TF sets that belong to complexes predicted by DACO shared common target genes, which is above mere chance for every *n*-tuple of TFs (Supplementary Fig. S4). This information could be used to build a gene regulatory network that includes potentially meaningful cooperative TFs. However, even more powerful predictions are feasible based on the DACO results because recruited regulatory proteins were predicted as well.

As an example for such an analysis, we have tried to characterize the different modes of action by which TF complexes may affect transcription. In general, TFs or regulatory proteins recruited by TFs either interact with the basal machinery or affect the chromatin structure and histone placement. Yeast only contains few proteins for histone acetylation and for the specific methylation of H3K4,36,79, whereas all modifications are basically associated with increased accessibility and therefore transcriptional upregulation (Millar and Grunstein, 2006). Few exceptions to this are known so far (Xin *et al.*, 2007). On the basis of this yeast-specific simplification, we compiled proteins associated with the following list of GO terms, namely direct contribution to a positive or negative transcriptional regulation (GO:0045944/GO:0000122), cofactor-mediated (GO:0003713/GO:0003714), and epigenetic contributions (GO:0004402, GO:0042054/GO:0004407, GO:0032452). Further, we determined chromatin remodeling proteins (GO:0006338) and proteins that belong to the basal machinery (GO:0016591). Single proteins that were annotated with both positive and negative influence in different contexts were excluded to prevent ambiguity (Supplementary Table S13).

Then, all protein assemblies predicted by DACO across all TF variants were checked for contradictory statements, e.g. whether they harbored proteins that are marked to positively influence transcription as well as proteins that are annotated with repressory function. Only 17% of the predicted complexes comprised conflicting proteins. Among the consistent candidates, 79% had at least one of the annotation as considered, and for 65% even the direction of the contribution to the regulation could be inferred. Interestingly, 3% of the TF combinations took part in assemblies with opposing regulatory effects. For example, the TF pair RFX1/REB1 occurred in 10 predicted candidates. Four of those were annotated with activating proteins, for two complexes repressory function was deduced and eight involved chromatin remodeling. All candidates were devoid of contradictory annotations. RFX1 is known to change its function by recruitment of the CYC8/TUP1 corepressor complex (Zhang and Reese, 2005). Both corepressor proteins were found in the predicted repressive candidates and were the decisive factors for the corresponding classification. Furthermore, REB1 is associated with nucleosome-depleted regions, supporting the recruitment of remodelers (Bai *et al.*, 2011).

While this simple categorization strategy seems to work well for yeast where it provided further mechanistic insights on such complexes, it will likely not be adequate for higher eukaryotes. Still, the potential information on the mode of action carried out by the recruited proteins is a huge benefit in comparison with the common information that is retrieved from the functions of TFs and target genes in gene regulatory networks. Future computational models of transcription may also consider the histone code, so that possible modifications could be inferred with an equivalent level of detail from recruited writer proteins.

3.3 Analysis of target gene coexpression

The complete set of predicted complex candidates provides a superordinate overview of putative regulatory players. To find causal links between TF tuples and their regulatory targets, we incorporated expression data and applied a strict association method. Based on the idea that genes that are regulated by the same control mechanism should exhibit a highly similar expression pattern in a certain condition, individual assemblies can be assigned to cellular states (Pilpel *et al.*, 2001; Wang and Zhang, 2006).

For all predicted pairs and higher-order tuples of TFs (290 in total), we analyzed the coexpression of their target genes within a time series of 32 timepoints for the yeast cell cycle by Chin *et al.* (2012). The aim of this was to determine significantly cooperative TF tuples that are assumed to be decisive regulatory drivers. Here, a tuple of n TFs is assumed to be decisive if the coexpression of the target genes significantly increases with the refinement induced by binding site constraints of an n-th TF. To quantify this measure, we used the methodology of Pilpel *et al.* (2001) with an expression coherence scoring (ECS) based on the correlation. To obtain a more accurate description of the binding mechanism, in addition to the unconstrained target genes shared by a set of TFs, also mutual targets were considered that allowed for colocalization of the TFs. These were defined as binding regions where all TFs showed pairwise distances in the range of –50 to +50 bp, which means overlaps and corecruitment were conceded. Also, targets of mediated cooperativity were defined as the genes that allowed for pairwise distances between 10 and 50 bp, and targets with supposed direct cooperativity were restricted to 0–10 bp.

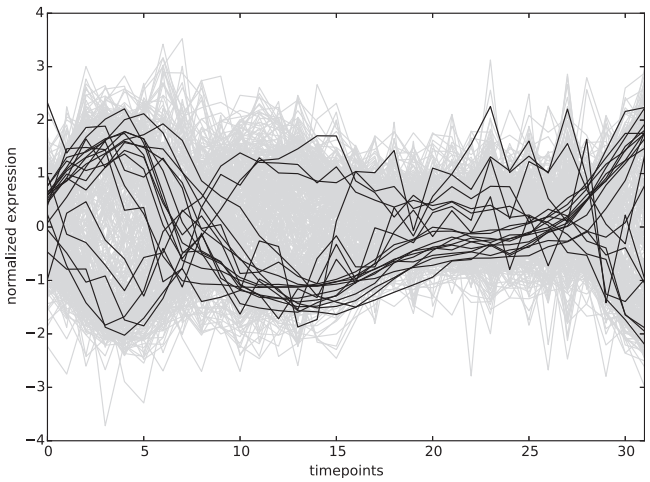

Fig. 5. Cell cycle expression profiles of all genes targeted by MET4 or MET32 (gray) are compared with the refined set of target genes where MET4 and MET32 bind as a colocalized complex with the two binding sites at pairwise distance between –50 and 50 bp (black). The increase in expression coherence between the targets of the individual TFs and the targets of a colocalized complex is highly significant (Table 2)

Seventeen higher-order TF combinations led to a significantly increased ECS in the context of the cell cycle ($P_{dECS} < 0.05$) for a certain binding mode. As an example, Figure 5 shows the complex-induced refinement in expression coherence among the target genes of the TF pair MET4/MET32. Subjected to GO term enrichment analysis, 76% of the corresponding target gene sets were significantly enriched with specific biological process annotations ($P < 0.05$, Bonferroni corrected). Table 2 summarizes the results of the evaluation. Not surprisingly, all significant tuples were associated with either cell cycle control itself or with metabolic processes that are in crosstalk with the cell cycle during normal growth. Most tuples are supported by literature evidence (Supplementary Table S14).

4 CONCLUSION

TF complexes are highly modular combinatorial assemblies and thus clearly different from large self-contained functional protein complexes. Our novel DACO approach was found to give superior results to established complex prediction programs for the

Table 2. The list of predicted TF combinations with significant increase of expression coherence (P_{dECS}) among their mutual targets comprised 15 pairs and two triples

TFs	P_{dECS}	Binding mode	Targets	Regulatory influence	GO process enrichment (P <0.05, Bonferroni corrected) in targets
MET4/MET32	0.0010	coloc.	19	+	Methionine metabolic process
TBP/HAP5	0.0335	med.	47	+	/
GLN3/DAL80	0.0009	med.	28	/	Allantoin catabolic process
DIG1/STE12/SWI6	0.0369	all	15	/	Fungal-type cell wall organization
FHL1/RAP1	0.0001	coloc.	116	+	rRNA transport
RPH1/GIS1	0.0001	med.	100	–	Hexose catabolic process
CBF1/MET32	0.0002	coloc.	33	o	Sulfate assimilation
DIG1/STE12	0.0003	med.	34	–	Response to pheromone
GCN4/RAP1	0.033	med.	62	+	/
MSN4/MSN2	0.0021	med.	105	+	Oligosaccharide biosynthetic process
DAL80/GZF3	0.0044	med.	20	–	Purine nucleobase metabolic process
SWI6/SWI4	0.0039	med.	53	+	Regulation of cyclin-dependent protein serine/threonine kinase activity
STB1/SWI6	0.0275	all	47	+	/
TBP/SWI6	0.0159	med.	14	+	/
GLN3/GZF3	0.0120	adj.	31	/	Allantoin catabolic process
MBP1/SWI6/SWI4	0.0307	med.	18	+	Regulation of cyclin-dependent protein serine/threonine kinase activity
MBP1/SWI6	0.0124	adj.	25	/	Cell cycle process

Note: Owing to the number of permutations of the test, the lowest possible value is $P_{dECS} = 10^{-4}$. The calculations were conducted for different conceivable modes of targeting (all shared target proteins, direct adjacency, mediated adjacency and colocalization) to have a detailed picture of the possible target–gene sets. Only the most enriched GO process term is shown for each target set. The inferred regulatory influence on the rate of transcription is abbreviated as follows: + (increase), – (decrease), o (no statement possible), / (conflicting annotations).

sophisticated task of predicting complexes involving TFs in the yeast *Saccharomyces cerevisiae*. In addition, we showed how the predictions can be used beneficially to identify individual complexes as regulatory drivers during a defined cellular state and condition.

ACKNOWLEDGEMENT

The authors thank Michael Hutter for a careful reading of the manuscript.

Funding: The work in the Helms group on gene regulatory networks is supported from the DFG through SFB1027. T.W. thanks the Saarbrücken Graduate School of Computer Science for a scholarship.

Conflict of interest: none declared.

REFERENCES

Aguilar,D. and Oliva,B. (2008) Topological comparison of methods for predicting transcriptional cooperativity in yeast. *BMC Genomics*, **9**, 137.

Aloy,P. and Russell,R.B. (2006) Structural systems biology: modelling protein interactions. *Nat. Rev. Mol. Cell Biol.*, **7**, 188–197.

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

Bader,G.D. and Hogue,C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.

Bai,L. *et al.* (2011) Multiple sequence-specific factors generate the nucleosome-depleted region on CLN2 promoter. *Mol. Cell*, **42**, 465–476.

Brohée,S. and van Helden,J. (2006) Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, **7**, 488.

Buchler,N.E. *et al.* (2003) On schemes of combinatorial transcription logic. *Proc. Natl Acad. Sci. USA*, **100**, 5136–5141.

Chang,D.T. *et al.* (2011) YPA: an integrated repository of promoter features in *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **39**, D647–D652.

Cherry,J.M. *et al.* (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.*, **40**, D700–D705.

Chin,S.L. *et al.* (2012) Dynamics of oscillatory phenotypes in Saccharomyces cerevisiae reveal a network of genome-wide transcriptional oscillators. *FEBS J.*, **279**, 1119–1130.

Diez,D. *et al.* (2014) Systematic identification of transcriptional regulatory modules from protein-protein interaction networks. *Nucleic Acids Res.*, **42**, e6.

Enright,A.J. *et al.* (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.

Friedel,C.C. *et al.* (2009) Bootstrapping the interactome: unsupervised identification of protein complexes in yeast. *J. Comput. Biol.*, **16**, 971–987.

Göke,J. *et al.* (2011) Combinatorial binding in human and mouse embryonic stem cells identifies conserved enhancers active in early embryonic development. *PLoS Comput. Biol.*, **7**, e1002304.

Han,J.D. *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**, 88–93.

He,Q. *et al.* (2011) High conservation of transcription factor binding and evidence for combinatorial regulation across six *Drosophila* species. *Nat. Genet.*, **43**, 414–420.

Hemberg,M. and Kreiman,G. (2011) Conservation of transcription factor binding events predicts gene expression across species. *Nucleic Acids Res.*, **39**, 7092–7102.

Hochedlinger,K. and Plath,K. (2009) Epigenetic reprogramming and induced pluripotency. *Development*, **136**, 509–523.

Huh,W.K. *et al.* (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691.

Hunter,S. *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.

Istrail,S. and Davidson,E.H. (2005) Logic functions of the genomic *cis*-regulatory code. *Proc. Natl Acad. Sci. USA*, **102**, 4954–4959.

Jansen,R. *et al.* (2002) Relating whole-genome expression data with protein-protein interactions. *Genome Res.*, **12**, 37–46.

Jung,S.H. *et al.* (2010) Protein complex prediction based on simultaneous protein interaction network. *Bioinformatics*, **26**, 385–391.

Junion,G. *et al.* (2012) A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell*, **148**, 473–486.

Kazemian,M. *et al.* (2013) Widespread evidence of cooperative DNA binding by transcription factors in *Drosophila* development. *Nucleic Acids Res.*, **41**, 8237–8252.

Keskin,O. and Nussinov,R. (2007) Similar binding sites and different partners: implications to shared proteins in cellular pathways. *Structure*, **15**, 341–354.

Kim,P.M. *et al.* (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*, **314**, 1938–1941.

Kim,Y. *et al.* (2012) IDDI: integrated domain-domain interaction and protein interaction analysis system. *Proteome Sci.*, **10** (**Suppl. 1**), S9.

Levine,M. and Tjian,R. (2003) Transcription regulation and animal diversity. *Nature*, **424**, 147–151.

Li,X. *et al.* (2010) Computational approaches for detecting protein complexes from protein interaction networks: a survey. *BMC Genomics*, **11** (**Suppl. 1**), S3.

Lopez,R. *et al.* (2003) Public services from the European Bioinformatics Institute. *Brief. Bioinform.*, **4**, 332–340.

Ma,W. *et al.* (2012) Protein complex prediction based on maximum matching with domain-domain interaction. *Biochim. Biophys. Acta*, **1824**, 1418–1424.

Manke,T. *et al.* (2003) Correlating protein-DNA and protein-protein interaction networks. *J. Mol. Biol.*, **333**, 75–85.

Mewes,H.W. *et al.* (2006) MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res.*, **34**, D169–D172.

Millar,C.B. and Grunstein,M. (2006) Genome-wide patterns of histone modifications in yeast. *Nat. Rev. Mol. Cell Biol.*, **7**, 657–666.

Nepusz,T. *et al.* (2012) Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods*, **9**, 471–472.

Ozawa,Y. *et al.* (2010) Protein complex prediction via verifying and reconstructing the topology of domain-domain interactions. *BMC Bioinformatics*, **11**, 350.

Pilpel,Y. *et al.* (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, **29**, 153–159.

Pu,S. *et al.* (2009) Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res.*, **37**, 825–831.

Punta,M. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.

Saito,R. *et al.* (2012) A travel guide to Cytoscape plugins. *Nat. Methods*, **9**, 1069–1076.

Siggers,T. and Gordan,R. (2014) Protein-DNA binding: complexities and multi-protein codes. *Nucleic Acids Res.*, **42**, 2099–2111.

Simon,I. *et al.* (2001) Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, **106**, 697–708.

Spitz,F. and Furlong,E.E. (2012) Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, **13**, 613–626.

The UniProt Consortium. (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **42**, D191–D198.

Tuncbag,N. *et al.* (2009) Towards inferring time dimensionality in protein-protein interaction networks by integrating structures: the p53 example. *Mol Biosyst.*, **5**, 1770–1778.

Tyagi,M. *et al.* (2009) Exploring functional roles of multibinding protein interfaces. *Protein Sci.*, **18**, 1674–1683.

von Mering,C. *et al.* (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399–403.

Wang,G. and Zhang,W. (2006) A steganalysis-based approach to comprehensive identification and characterization of functional regulatory elements. *Genome Biol.*, **7**, R49.

Wilson,N.K. *et al.* (2010) Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell*, **7**, 532–544.

Xin,X. *et al.* (2007) Regulation of the HAP1 gene involves positive actions of histone deacetylases. *Biochem. Biophys. Res. Commun.*, **362**, 120–125.

Yellaboina,S. *et al.* (2011) DOMINE: a comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Res.*, **39**, D730–D735.

Zhang,B. *et al.* (2008) From pull-down data to protein interaction networks and complexes with biological relevance. *Bioinformatics*, **24**, 979–986.

Zhang,Q.C. *et al.* (2013) PrePPI: a structure-informed database of protein-protein interactions. *Nucleic Acids Res.*, **41**, D828–D833.

Zhang,Z. and Reese,J.C. (2005) Molecular genetic analysis of the yeast repressor Rfx1/Crt1 reveals a novel two-step regulatory mechanism. *Mol. Cell. Biol.*, **25**, 7399–7411.