YEASTRACT: a database of transcription regulatory associations in *Saccharomyces cerevisiae*

Pedro Monteiro[1], Miguel C. Teixeira[2], Pooja Jain[1],
Sandra Tenreiro[2], Alexandra R. Fernandes[2], Nuno Mira[2], Marta Alenquer[2],
Ana T. Freitas[1], Arlindo L. Oliveira[1], and Isabel Sá-Correia[2]

[1] INESC-ID/IST, Lisbon Portugal
[2] BSRG, CEBQ, IST, Lisbon, Portugal

Abstract. We present the YEASTRACT (Yeast Search for Transcriptional Regulators And Consensus Tracking; www.yeastract.com) database. This database is a repository of more than 9000 regulatory associations between genes and transcription factors in *Saccharomyces cerevisiae*, based on more than 900 bibliographic references. It also includes the description of 242 specific DNA binding sites for 102 characterized transcription factors. Further information about each yeast gene included in the database was extracted from a number of different sources, and combined in order to make available a number of queries related with regulatory processes in Yeast.
All the information in YEASTRACT will be updated regularly to match the latest data from SGD, GO consortium and recent literature on yeast regulatory networks.
Future releases of the database will include additional computational tools to support researchers in the process of identification of transcription regulatory associations.

Keywords: Consensus, Gene Transcription Regulation, Transcriptional regulators, Transcription factor binding sites, Saccharomyces cerevisiae, Database

1 Introduction

The model eukaryote Saccharomyces cerevisiae, with the genome sequence available since 1996 [3], plays an essential role in our efforts to understand the complex biological networks that control life processes.

Since the mapping and reporting of the genome sequence of the Yeast *S. cerevisiae*, a number of computational methods and tools have become available to support research related with this organism. Most significant for the Yeast community, the SGD database [1] makes available extensive information on a number of aspects of genomic information of Yeast. Other databases that include relevant information and computational tools that can be used to study regulatory mechanisms in Yeast like RSAT are also available [4].

However, none of the existing databases gives adequate and specific support to the complex process of integrating evidence obtained from sequence, functional and expression analysis, in particular those emerging from global expression analysis, in order to identify or to guide the prediction of genetic regulation mechanisms.

YEASTRACT proposes to fill this gap, by making publicly available a database that contains curated, up to date, information about regulatory associations in Yeast.

2 Database content

The YEASTRACT database contains information about regulatory associations in yeast. Given that the three principal entities involved in gene regulation are the concept of gene, protein and binding site, the internal structure of the database is organized around these three concepts.

Figure 1 describes the major entities in the database, and their relationships.

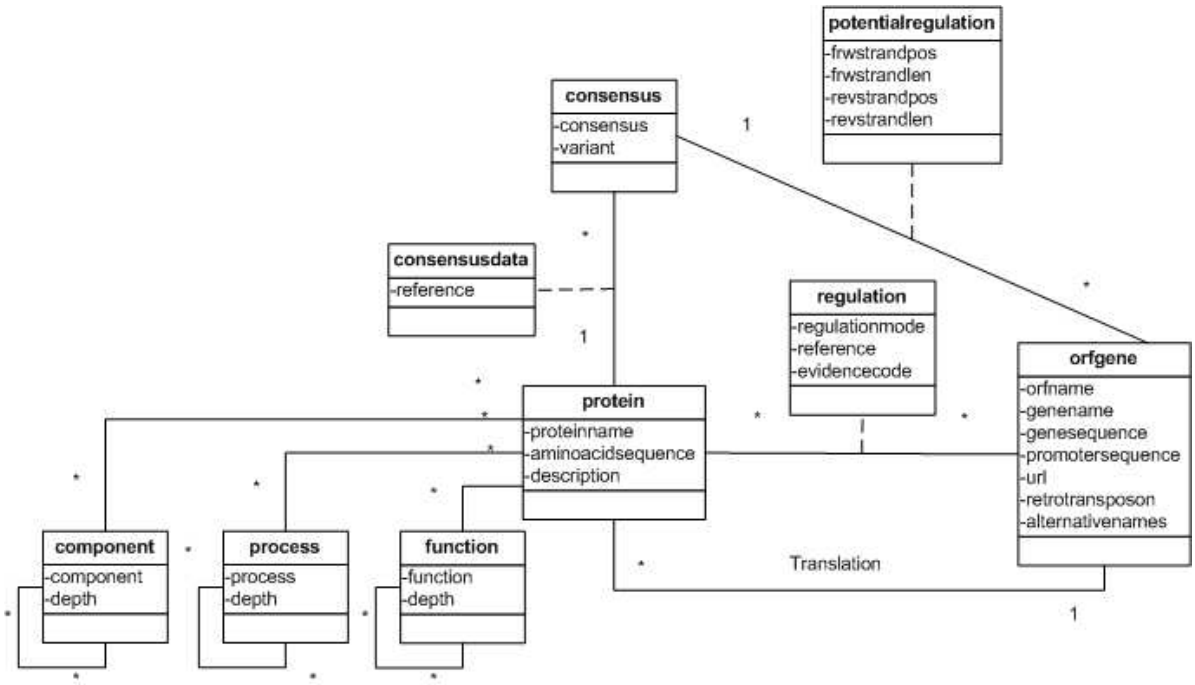

Fig. 1. Main entities in the database model and their relationships

The concept of *ORF/gene* refers to a DNA region limited by start and stop codons. If the gene that corresponds to a given ORF is characterized, the gene name is stored as one of the attributes. A number of other attributes is also stored for each entity ORF/Gene. *Proteins* are also fundamental concepts and are stored as separate entities, with specific attributes. Finally, the concept of *consensus* models the binding sites that mediate the process of regulation of genes by transcription factors.

The principal relation between these three concepts is the *regulation* relation. This relation documents the associations between transcription factors and target genes, and can be of two types: documented and potential.

A documented association between a transcription factor and a target gene is supported by published data showing at least one of the following experimental evidences:

- Change in the expression of the target gene due to the deletion (or mutation) of the transcription factor encoding gene; these evidences may come from detailed gene by gene analysis or genome-wide expression analysis
- Binding of the transcription factor to the promoter region of the target gene, as supported by band-shift, foot-printing or Chromatine ImmunoPrecipitation (ChIP) assays. The user should check the corresponding references, provided in the database, to become fully aware of the nature of the evidences.

Each documented association is annotated with the reference or references that provided the information.

A potential association between a transcription factor and a target gene is based on the occurrence of the transcription factor binding site in the promoter of the target gene. The binding sites that were considered for each transcription factor are supported by published footprinting or ChIP experiments. As above, references are provided to be checked by the user, if desired.

Figure 1 also shows, on the top left, some auxiliary concepts that correspond to tables used to track and log usage of the database.

The database presently contains more than 9000 regulatory associations between genes and transcription factors, based on more than 900 bibliographic references. Each regulation has been annotated by hand, after examination of the relevant references. The database also contains 242 specific DNA binding sites for 102 characterized transcription factors, documented in the relevant literature.

3 User interface and queries

YEASTRACT was made available on the web after extensive internal testing. The interface was developed by a group that included software engineers and biologists, with the objective of improving its usefulness and usability. All pages have a context dependent help, as well as a context dependent example of utilization. A complete tutorial on the use of the system is also available.

3.1 Available queries

YEASTRACT makes available the information stored in the database through a set of queries and a number of additional utilities.

The major queries available in the present version are:

- Search transcription factors by regulated genes or by keywords
- Search genes regulated by specific transcription factors
- Group Genes by transcription factor
- Search by DNA motif
- Search regulatory associations between transcription factors and genes

The *search transcription factors by regulated genes* query allows the user to identify documented and/or potential regulators of genes present in the given list of genes. Documented and potential targets of each transcription factor, which are present in the given gene list, are displayed in a table. This search rejects the transcription factors for which no documented or potential regulatory associations with any of the input genes exists. By default regulatory associations are searched for input transcription factors against input genes, but other options are available to the user.

Another facility provided by this query is the search for transcription factors that match a specific set of keywords, provided by the user. This query allows the user to search for transcription factors by keywords, found to occur in their description, as extracted from the Saccharomyces Genome Database [1].

The *search genes regulated by specific transcription factors* allows the user to search for genes regulated by the given transcription factors. The user may search for the genes documented as being regulated by specific transcription factors or for genes potentially regulated, based on the

existence of the transcription factor binding site in their promoter region. It is also possible to combine the results from the above two searches.

The *group genes by transcription factor* query allows the user to group a given gene list (for instance a set of co-activated genes coming out of a microarray experiment) according to the transcription factors which are their documented or potential regulators. The transcription factors considered during this query are either a given list of transcription factors or all the transcription factors in the database. For each established regulon, a percentage value representing the proportion of genes regulated by each transcription factor, is presented. This value is calculated relative to a) the total number of genes in the given list; this may indicate the transcription factors involved in the regulation of the referred gene list or b) the number of genes, in the whole yeast genome, documented as being regulated by the same transcription factor; this may indicate the transcription factor networks predominantly involved in the regulation of the referred gene list.

The *search by DNA motif* query allows the user to search one or more DNA motifs in the promoter region of one or more genes or within the described transcription factor binding sites. The first possibility, the search for DNA motifs within the promoter region of genes searches for the input DNA motifs in the promoter region of one or more genes or within the promoter region of all the genes in the database. There is an imposed minimum length for motifs used in this query. The second possibility searches for a match between a single DNA motif with described transcription factors binding sites. The result of this search is a list of transcription factors binding sites where there exists a match with the given DNA motif. This search allows the user to check whether a newly identified DNA motif corresponds to a previously described transcription factor binding site. The search also accepts DNA motifs with ambiguous bases (IUPAC code).

Finally, the *search regulatory associations between transcription factors and genes* query allows the user to identify documented as well as potential regulatory associations between input transcription factors and genes. Documented and potential targets of each transcription factor, which are present in the given gene list are displayed in a table. This search rejects the transcription factors for which no documented or potential regulatory associations with any of the input genes exists. By default regulatory associations are searched for input transcription factors against input genes, but other options are available to the user

3.2 Additional utilities

A number of additional utilities is also available in YEASTRACT. These utilities can be used by themselves, or coupled with the previously described main queries. Some of these utilities use data from the Gene Ontology consortium [2]. The most relevant utilities provided are:

- Group genes by gene ontology terms
- Group regulations by gene ontology terms
- Transform a list of ORFs into a list of genes

The *group genes by gene ontology terms* groups a given set of genes according to the gene ontology terms assigned to them by SGD. Grouping can be done by either of the three gene ontologies, namely: Biological Process, Molecular Function and Cellular Component. The specificity of grouping can be enhanced by specifying the level of the gene ontology terms in their respective hierarchy that increases from 2 to 6.

The *group regulations by gene ontology terms* utility supplements the search for documented and potential regulatory associations between input lists of transcription factors and regulated genes by grouping regulatory associations by gene ontology terms. The grouping is anchored at the gene ontology terms associated to the transcription factor. Furthermore, the user can select either the Biological Function or the Molecular Process ontology.

Finally, the *transform a list of ORFs into a list of genes* utility is a general purpose utility made available to simplify the process of querying databases and systems that accept only one of the names. When an ORF has no attributed gene name, the ORF name appears in the gene name list.

4 Typical applications

Although the database can be used in large number of ways to process data from different sources, and with different objectives, three major processes are directly supported by the existing query facilities.

The first process consists in the identification of documented and potential regulatory associations for an ORF/Gene.

The second process is related with the grouping of genes with identified common expression profiles (e.g., from microarray data) based on their regulatory associations.

The third process is related with the identification of putative binding sites, performed by searching for a DNA motif within known transcription factor binding sites and promoter regions.

The tutorial provided in the database presents three case studies exemplifying the use of different query options and utilities to attain these three main objectives

5 Acknowledgments

The information about Yeast genes other than documented regulations, potential regulations and the transcription binding sites contained in YEASTRACT was gathered from Saccharomyces Genome Database (SGD), Gene Ontology (GO) Consortium and Regulatory Sequence Analysis Tools (RSAT).

We acknowledge the RSAT team for providing free access to the different functionalities of RSAT to academic users. We are also grateful to colleagues and friends from the Yeast community for their encouragement and suggestions. A very special thanks is due to André Goffeau.

This work was partially supported by project POSI/EIA/57398/2004, *DBYeast, A Framework for the development of algorithms to the analysis and identification of gene regulatory networks*, financed by FCT and the POSI program.

References

1. J. Michael Cherry, Caroline Adler, Catherine A. Ball, Stephen A. Chervitz, Selina S. Dwight, Erich T. Hester, Yankai Jia, Gail Juvik, TaiYun Roe, Mark Schroeder, Shuai Weng, and David Botstein. SGD: Saccharomyces genome database. *Nucleic Acids Research*, 26(1):73–79, 1998.
2. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.
3. A. Goffeau, B.G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J.D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S.G. Oliver. Life with 6000 genes. *Science*, 274(546):563–567, 1996.
4. J. van Helden. Regulatory sequence analysis tools. *Nucleic Acids Research*, 31(13):3593–3596, 2003.