UNIVERSIDADE DE LISBOA
Faculdade de Ciências
Departamento de Informática

GREAT: GENE REGULATION EVALUATION TOOL

Cátia Maria Machado

Mestrado em Tecnologias de Informação aplicadas às Ciências Biológicas e Médicas

2009

UNIVERSIDADE DE LISBOA

Faculdade de Ciências
Departamento de Informática

GREAT: GENE REGULATION EVALUATION TOOL

Cátia Maria Machado

Trabalho orientado pelo Prof. Dr. Francisco José Moreira Couto

Mestrado em Tecnologias de Informação aplicadas às Ciências Biológicas e Médicas

2009

Resumo

A correcta compreensão de como funcionam os sistemas biológicos depende do estudo dos mecanismos que regulam a expressão genética. Estes mecanismos controlam em que momento e durante quanto tempo é utilizada a informação codificada num gene, e podem actuar em diversas etapas do processo de expressão genética. No presente trabalho, a etapa em análise é a transcrição, na qual a sequência de ADN de um gene é transformada numa sequência de ARN, que posteriormente dará origem a uma proteína.

A regulação da transcrição centra-se na acção de uma classe de proteínas reguladoras denominadas factores de transcrição. Estes ligam-se à cadeia de ADN na região próxima do início de um gene (a região promotora), potenciando ou inibindo a ligação da proteína responsável pelo processo de transcrição.

Os factores de transcrição têm especificidade para pequenas sequências de ADN (denominados motivos de ligação) que estão presentes nas regiões promotoras dos genes que regulam. Um gene pode ser regulado por diferentes factores de transcrição; um factor de transcrição pode regular diferentes genes; e dois factores de transcrição podem ter motivos de ligação iguais.

A regulação dos genes que codificam factores de transcrição é ela própria regulada, podendo sê-lo por uma série de mecanismos que incluem a interacção com outros factores de transcrição.

O conhecimento de como genes e proteínas interagem entre si permite a criação de modelos que representam o modo como o sistema em questão (seja um processo biológico ou uma célula) se comporta. Estes modelos podem ser representados como redes de regulação genética, que embora possam diferir estruturalmente, os seus componentes elementares podem ser descritos da seguinte forma: os vértices representam genes (ou as proteínas codificadas) e as arestas representam reacções moleculares individuais, como as interacções entre proteínas através das quais os produtos de um gene afectam os de outro.

A representação de regulações genéticas em redes de regulação genética promove, entre outros aspectos, a descoberta de grupos de genes que, sendo co-regulados, participam no mesmo processo biológico.

Tal como referido anteriormente, os factores de transcrição podem ser regulados por outros factores de transcrição, o que significa que existem dois tipos de regulações: directas e indirectas. Regulações directas dizem respeito a pares gene-factor de

transcrição em que a expressão do gene é regulada pelo factor de transcrição considerado no par; regulações indirectas dizem respeito a pares em que a expressão do gene é regulada por um factor de transcrição cuja expressão é regulada pelo factor de transcrição considerado no par.

Existem dois tipos de métodos experimentais que permitem a identificação de regulações genéticas: métodos directos, que identificam regulações directas; métodos indirectos, identificam regulações mas sem ser possível diferenciar entre directas e indirectas. Os métodos directos avaliam a ligação física do factor de transcrição ao gene, enquanto os métodos indirectos avaliam a existência de alterações nos padrões de expressão dos genes devido à influência dos factores de transcrição (isto é, se a acção de um determinado factor de transcrição se deixar de sentir, quais os genes cuja transcrição sofrerá alterações, e com que intensidade).

Dos quatro métodos descritos em seguida, os dois primeiros são directos e os dois últimos indirectos:

- *Chip* (imunoprecipitação de cromatina) – esta técnica é utilizada na investigação de interacções *in vivo* entre DNA e proteínas [1,2].
- *CHIP-chip* – esta técnica consiste numa adaptação da anterior, sendo realizada à escala genómica: um *microarray* representativo do genoma completo de um organismo é exposto a um dado FT, permitindo a identificação de todos os genes que este regula [3].
- *Microarrays* – a utilização de *microarrays* permite a avaliação de alterações de expressão genética em grande escala, considerando o genoma completo de um organismo ou apenas uma via metabólica [4].
- Proteómica – esta abordagem inclui diversos métodos que permitem a identificação dos genes regulados por um determinado factor de transcrição através do estudo do nível de expressão das proteínas codificadas pelos genes [5].

O conhecimento existente sobre regulações genéticas encontra-se disponível essencialmente na literatura. Embora actualmente exista um número elevado de bases de dados biológicas públicas, a grande maioria contém dados sobre entidades biológicas mas não sobre regulações genéticas de forma explícita.

Com o objectivo de colocar à disposição da comunidade científica dados existentes sobre regulações genéticas em *Saccharomyces cerevisiae*, foi criada uma base de dados portuguesa, denominada Yeastract, mantida por curação manual de literatura científica.

Devido à crescente quantidade de artigos publicados actualmente, é de extrema importância o desenvolvimento de ferramentas automáticas que auxiliem o processo de curação manual. No caso concreto da Yeastract, surgiu a necessidade de criar uma ferramenta que auxiliasse o processo de identificação de artigos científicos que descrevam regulações genéticas em *S. cerevisiae*. Esta ferramenta é composta por dois componentes: um primeiro que identifica factores de transcrição nos resumos dos artigos e que verifica se os resumos contêm descrições de regulações genéticas; um segundo que avalia se as regulações hipotéticas que o artigo contém correspondem a regulações válidas do ponto de vista biológico. Este segundo componente foi denominado GREAT (*Gene Regulation EvAluation Tool*) e constitui o objectivo do meu trabalho.

A ferramenta que desenvolvi recebe como *input* uma lista de artigos em cujos resumos foram identificados factores de transcrição e, na validação das regulações, explora dados obtidos exclusivamente de bases de dados biológicas de acesso público. Esses dados são utilizados na avaliação dos seguintes aspectos: participação de um gene e de um factor de transcrição no mesmo processo biológico; existência do motivo de ligação do factor de transcrição na região promotora do gene; método experimental com que a regulação foi identificada. O resultado de cada um destes aspectos é utilizado por um método de aprendizagem automática, árvores de regressão ou árvores modelo, para o cálculo de um *score* de confiança, a atribuir a cada potencial regulação. Artigos que contenham regulações com *scores* elevados serão curados manualmente para extracção das regulações genéticas.

Foi implementado com sucesso um primeiro protótipo do GREAT. No entanto, do ponto de vista biológico, os resultados obtidos não foram satisfatórios, pelo que se realizou uma análise detalhada dos dados utilizados. Esta análise revelou questões importantes, essencialmente relacionadas com a insuficiência de dados disponíveis, e permitiu a identificação de medidas que poderão ser implementadas no actual protótipo para a resolução dos problemas encontrados.

Palavras-Chave: Regulações Genéticas, Regulação da Transcrição, Bases de Dados Biológicos Públicas, *Gene Ontology*

Abstract

The understanding of biological systems is dependent on the study of the mechanisms that regulate gene expression. These mechanisms control when and for how long the information coded in a gene is used, and can act on several of the steps in the gene expression process. In the present work, the step of interest is the transcription, where the DNA sequence of a gene is transformed into an RNA sequence, which will later be used to synthesise a protein.

The knowledge about gene regulations is mainly available in the literature. Although there are currently multiple public biological databases, the majority of those contain data on biological entities but not explicitly on gene regulations.

In order to provide the scientific community with data on *Saccharomyces cerevisiae* transcription regulations, a Portuguese public repository maintained by manual curation of scientific literature, named Yeastract, was created.

Due to the increasing amount of papers published nowadays, the development of automatic tools that can help the curation process is of great importance. In the specific case of Yeastract, a tool was needed to help in the identification of papers describing gene regulations of *S. cerevisiae*. This tool was created with two components: one that identifies transcription factors in the papers' abstracts and verifies if they describe gene regulations; the other that evaluates if the hypothetical regulations the paper contains correspond to valid regulations from a biological point of view. This second component was named GREAT, Gene Regulation EvAluation Tool, and is the goal of my work.

The tool I developed uses data obtained exclusively from public biological databases to validate the regulations. That data is used in the evaluation of three aspects: the participation of a gene and a transcription factor in the same biological process; the existence of the transcription factor binding motif in the gene promoter region; the experimental method with which the regulation was identified. The output of these features is used by a machine learning method, either regression or model trees, to calculate a confidence score to attribute to each putative gene regulation. Papers containing regulations with high scores will be manually curated to extract the gene regulations.

Although a first prototype of GREAT was implemented, from a biological point of view the results obtained are unsatisfactory. This prompted a detailed analysis of the data used, which uncovered important questions that need to be addressed in order to further improve this tool.

Keywords: Gene Regulation, Transcription Regulation, Public Biological Databases, Gene Ontology

Acknowledgments

I would like to thank my advisor, Dr. Francisco Couto, for his shared knowledge and support throughout this last year of my life, during which not only this work was developed but also my first steps in bioinformatics were taken.

I thank Daniel Faria for all the conversations about my personal and professional self during the years we have had the opportunity to share.

I'm also grateful to Hugo Bastos for helping me in the beginning of this work, to Cátia Pesquita for her appropriate comments and suggestions, and to Ana Teixeira for her company in the so useful daily breaks. The company of these people as well of Tiago Grego, Filipe Lopes and Francisco Lopez-Pellicer provided a very agreeable working environment.

Thank you.

Lisboa, July 29, 2009
Cátia Maria Machado

Contents

List of Tables

Acronyms

DNA – Desoxirribonucleic Acid

GO – Gene Ontology Database

GOA – Gene Ontology Annotation Database

IC – Information Content

ML – Machine Learning

NLP – Natural Language Processing

RMSE – Root Mean Squared Error

RNA – Ribonucleic Acid

SGD – Saccharomyces Genome Database

SVM – Support Vector Machine

TF – Transcription Factor

Chapter 1

Introduction

The understanding of biological systems is dependent on the study of the mechanisms that regulate gene expression. These mechanisms control when and for how long the information coded in a gene is used, and can act on several of the steps in the gene expression process. In the present work, the step of interest is the transcription, where the DNA sequence of a gene is transformed into an RNA sequence, which will later be used to synthesise a protein.

The regulation of the transcription is centered on the activity of regulatory proteins, called transcription factors, which bind to a region of the DNA sequence near the origin of the gene (the promoter region), enabling or inhibiting the binding of the protein responsible for the transcription process.

Transcription factors recognize specific DNA motifs, present in the promoter region of the genes, thus identifying their targets. One gene can be regulated by several transcription factors; a transcription factor can regulate several genes; and two transcription factors can have the same binding motif.

Since transcription factors are themselves encoded by genes, their expression is also subject to regulation including the interaction with other transcription factors.

The knowledge of how genes and gene products interact with each other enables the creation of models that represent how the system in question (a specific biological process or a cell as a whole) behaves. These models can be represented as gene regulatory networks, which can vary greatly in structure but whose elemental components can be described as follows: the nodes represent genes (or their products) and the edges represent individual molecular reactions, such as protein interactions where the products of one gene affect those of another.

Amongst other aspects, the representation of gene regulations in gene regulatory networks promotes the uncovering of groups of genes that, being co-regulated, participate in the same biological process.

As referred before, transcription factors themselves can be regulated by other transcription factors, which means that there are two types of regulations: direct and indirect. Direct regulations refer to gene-transcription factor pairs whose gene's expression is regulated by the binding of the transcription factor considered in the pair; indirect regulations refer to pairs whose gene is bound by a transcription factor that is itself regulated by the transcription factor considered in the pair.

There are two types of experimental methods that allow the identification of gene regulations: direct methods, with which direct regulations are identified; indirect methods, with which regulations are identified but without the possibility to differentiate direct from indirect. Direct methods verify the physical binding of the transcription factor to the gene promoter region, while indirect methods identify changes in the expression patterns of genes due to the influence of the transcription factors (i.e. if the action of a given transcription factor is somehow hampered, which genes will have their transcription affected, and how strongly).

From the four methods described next, the first two are direct and the last two indirect:

- ChIP (Chromatin ImmunoPrecipitation) – this technique is used to investigate interactions between DNA and proteins *in vivo* (such as transcription factors) [1,2].
- ChIP-chip – this technique is an adaptation of ChIP to a genomic-wide scale: a microarray representative of an organism whole genome is incubated with a given TF, allowing for the identification of all of its gene targets [3].
- Microarrays – the utilization of microarrays enables the evaluation of gene expression changes in a large-scale, either whole genome or just a pathway [4].
- Proteomics – this approach encloses several methods, which allow the identification of the genes regulated by a given transcription factor through the study of the expression levels of the proteins they encode [5].

The knowledge concerning gene regulations is available mainly from the literature, currently the preferred mean of scientific dissemination. Although a great amount of public biological databases exist nowadays, the great majority contain data on biological entities but not explicitly on gene regulations.

In order to provide the scientific community with data on *Saccharomyces cerevisiae* transcription regulations, Yeastract [6] was created. This a Portuguese public repository maintained by literature manual curation, providing not only the data but also a set of bioinformatics tools to explore it.

1.1 Motivation

Since the scientific literature is growing at an ever increasing rate [7], its manual curation has become unfeasible. As such, the development of an automatic tool that can identify papers containing the sought information, in this case, *S. cerevisiae* gene regulations, is of great importance.

Although text mining tools can be used to identify regulations, their development still depends on the domain knowledge provided by humans which is often difficult to translate into machine-usable information [8].

1.2 Objectives

The goal of my work is to develop a tool, named GREAT (Gene Regulation Evaluation tool) that, given a list of abstracts and the transcription factors (TFs) identified in them, calculates a confidence score for each gene-TF pair in the paper that states if the pair corresponds to a true gene regulation. The papers containing pairs with high scores will be manually curated for the extraction of the gene regulations and subsequent storage in Yeastract database.

The list of abstracts and TFs is obtained from the output of a text mining tool that verifies the existence of TFs in the papers abstracts, and if the sentences in which the TFs are found may describe a gene regulation.

The external data used in GREAT is obtained from scientific literature and public biological databases.

These objectives are integrated in the project “ARN – Algorithms for the Identification of Genetic Regulatory Networks” (PTDC/EIA/67722/2006).

1.3 Methodology

The input of GREAT is a list of PubMed Identifiers (PubMed Id) of papers whose abstracts were identified as containing at least one TF (automatic identification with a text mining tool).

For each paper, GREAT obtains the genes referenced therein from the Saccharomyces Genome Database (SGD) and pairs up all these genes with all the TFs identified in the text to represent all possible gene regulations that may be described in the paper.

For each gene-TF pair obtained, a confidence score attribution is performed through the evaluation of the following three features:

- Biological Potential – if a gene is regulated by a TF, they have to participate in the same biological process (data source: Gene Ontology Database (GO)).
- Physical Potential – if a gene and a TF can physically bind, then it's possible that the gene is regulated by that TF (data source: Yeastract database).
- Experimental Evidence – if a direct method was used to evaluate the gene regulation, then the gene and the TF can physically bind (data source: SGD).

These three features are used by a machine learning method to calculate, for each gene-TF pair, a confidence score in the interval [0,1] – the closer to 1, the higher the confidence that the pair represents a regulation.

1.4 Results

A first prototype of GREAT was implemented. This prototype explores the data obtained from public biological databases, and attributes a score to each gene-TF pair identified, indicative of how likely that pair represents a gene transcription regulation.

From the biological point of view, the results obtained with the training/evaluation data were unsatisfactory. A detailed analysis of the data revealed: the existence of problems related to, among other aspects, the insufficient availability of data (and/or data sources); directions that can be followed in order to solve the problems encountered.

1.5 Document Organisation

This document is organized in the following manner:

- Chapter 2 – contains a brief explanation on the methods commonly used in the identification of gene transcription regulations in scientific literature.
- Chapter 3 – describes Yeastract database and all the external databases used in the implementation of GREAT or identified as potentially useful for that purpose.
- Chapter 4 – contains the details of the design and implementation of GREAT.
- Chapter 5 – describes the results obtained and their analysis.
- Chapter 6 – analyses the fulfilment of the proposed objectives and proposes some future work directives.
- Attachment – contains examples of hypothetical false negative pairs.

Chapter 2

Identification of gene transcription regulations in the literature

The identification and extraction of gene regulations from text is a difficult task due to the intrinsic complexity of both the natural language and the domain terminology. A particular piece of information can be expressed in more than one sentence in a document (or abstract), sometimes implicitly, and using different synonymous expressions. Furthermore, in the scientific literature, particularly in Biology, a great amount of domain-specific terminology is used, with new terms and variations in constant formation.

Techniques provided by natural language processing (NLP) are used to deal with human language, exploiting its multi-level regularities and constraints. Some of the levels considered include the following:

- Words – the basic building block of language, a word comprehends a root and possibly prefixes and suffixes.
- Syntax (or grammar) – controls how words are grouped into meaningful sentences, and its analysis can involve the tagging of each word to distinguish nouns from verbs, for instance.
- Semantics – semantic relations capture the meaning of the words, independently of the syntax and the actual words used [7].

Dictionaries and ontologies provide an important assistance in the interpretation of scientific literature. Lexical databases, like WordNet [9,10], provide a more general knowledge of the English language (in which the majority of the scientific papers are written), and biomedical ontologies, like GO [11], provide domain-specific knowledge.

Another way to insert domain knowledge into a system is through the identification of specific words that are expected to be found, like transcription factors in gene transcription regulations.

There are two approaches normally followed for the extraction of binary relations from biomedical text: symbolic pattern-based systems (rule systems) and feature-based statistical machine learning (ML) systems. Specifically for the extraction of gene transcription regulations, both type of systems need to perform the following steps:

- Identification of pairs of gene references as the arguments of the relation (entity recognition).
- Identification of the roles of the arguments in the relation (the regulator and the regulated).
- Decision whether the entity pair constitutes a relation [12].

The referred types of systems employ NLP techniques to various extends, whether just for simpler tasks as sentence splitting and tokenization, or for the implementation of any of the steps described above.

Both rule-based and ML-based approaches present advantages and disadvantages: while the development of rules allows an easier incorporation of semantic and biological constraints [13], the fact that they are fine-tuned for a specific application may render them less easily adaptable to changes in the application area; in the case of ML approaches, since they are trained with annotated corpora (either automatically or manually), the adaptation to changes is more easily accomplished [14], but the enforcement of constraints may be restricted (if they are not present in the training corpora, the system will not learn them).

2.1 State-of-the-art systems

Hahn *et al.* [14] describe one rule-based system and one ML-based. Both aim at the identification of gene transcription regulations from full texts, using the RegulonDB [15] as a gold standard: regulations identified by the systems and present in the database are considered true positives, regulations identified by the systems and not present in the database are considered false positives.

Regarding the steps that these systems need to perform in order to extract the gene regulations, the rule-based system implements them in the following manner:

- Entity recognition – the system bases the identification of names on a list of possible names obtained from RegulonDB.
- Relation identification – the system analyses the syntactic and semantic structures of the sentences through the utilization of patterns manually created for keywords related to gene regulation.
- Relation evaluation – this step is based on the manual creation of inference rules that reflect the knowledge of the gene regulation domain and that, when applied to the patterns previously referred, allow the inference of implicit meanings in the text.

In the case of the ML-based system, the referred steps are implemented in the following manner:

- Entity recognition – the system uses a ML-based name tagger trained with publicly available corpora.
- Relation identification and evaluation – the system employs Maximum Entropy models [16] considering text features, namely word features (the words before, after and between the recognised entities) and entity features (account for combinations of entity types).

Table 1 contains the precision and recall obtained with both rule-based and ML-based systems. The pairs identified as false positives in the ML-based system were analysed in detail and 21% of them correspond to true regulations that are not present in the RegulonDB.

Table 1. Evaluation statistics of the gene regulations' identification systems described by Hahn [14] and Saric [13].

Evaluation Statistics	Hahn		Saric
	Rule-based system	ML-based system	
Precision (%)	53	54	83-90[1]
Recall (%)	5.6	10	20

[1] – Variations dependent on the biological organism considered.

Saric *et al.* [13] describe a rule-based system whose purpose is to extract from biological abstracts information on which proteins are responsible for regulating the expression of genes, independently of the organism.

The accuracy of the relations was evaluated at the semantic rather than at the grammatical level: regulations identified by the system were considered true positives if they extracted the correct biological conclusion, independently of the analysis of the sentence from a linguistic point of view.

The main steps performed by the system were implemented as follows:

- Entity recognition – this step is performed using cascades of finite state rules [17]. The system uses a dictionary of synonymous names and identifiers of six eukaryotic model organisms, extended to include different orthographic variants of each name.
- Relation identification – the system also recognizes verbs – of activation, repression, etc – to improve this specific step. The combination of syntactic and semantic properties of the relevant verbs allows their mapping to the relations recognized (up, down and unspecified regulation of expression).
- Relation evaluation – this step is performed manually for all regulations extracted from the evaluation corpus using the TIGER Search visualization tool [18].

The precision and recall statistics obtained with this system are present in Table 1.

2.2 GREAT

GREAT is part of a two component system whose purpose is the identification of gene transcription regulations in abstracts, specifically for the model organism *S. cerevisiae*. The Yeastract database is used as a gold standard, with regulations identified by the system and present in the database being considered true positives and those not present in the database being considered false positives. Like the systems described above, this one also includes the following steps:

- Entity recognition – TF names are identified by the first component of the system, using an ML-approach (88% of precision and 90% of recall); gene names are obtained by GREAT from a public database that contains the list of genes referenced in each paper. For the identification of the TFs the system uses a dictionary of names obtained from Yeastract and SGD.
- Relation identification – all potential regulation relations present in each paper are considered through the pairwise combination of all TFs with all genes.

- Relation evaluation – this step is performed by GREAT using a ML algorithm (either regression or model trees) that calculates a confidence score to attribute to each relation. The algorithm combines the output of three features: physical potential, biological potential and experimental evidence.

Among other aspects, text mining approaches are highly dependent on the efficient identification of the biological entities and on their correct semantic tagging. The gene regulations' identification system in which GREAT is included only uses text-mining in the identification of the TFs. All steps performed by GREAT take advantage of data already curated and publicly available.

Since this system was designed specifically to help the manual curation process of Yeastract, some of the data sources used by GREAT are specific to *S. cerevisiae*. Nevertheless, the principles in which the regulations' evaluation features were constructed upon are not species-specific.

In GREAT, the domain knowledge is imbued in the definition and implementation of the evaluation features, and is straightforward in terms of machine utilization since the data does not need to be "interpreted" but only collected as specified.

To the best of my knowledge, there are no gene regulations' identification systems that use more than the text itself in the identification of the relations' entities, or that use features based on data from databases instead of text features like GREAT does.

Chapter 3

Yeastract and Related Resources

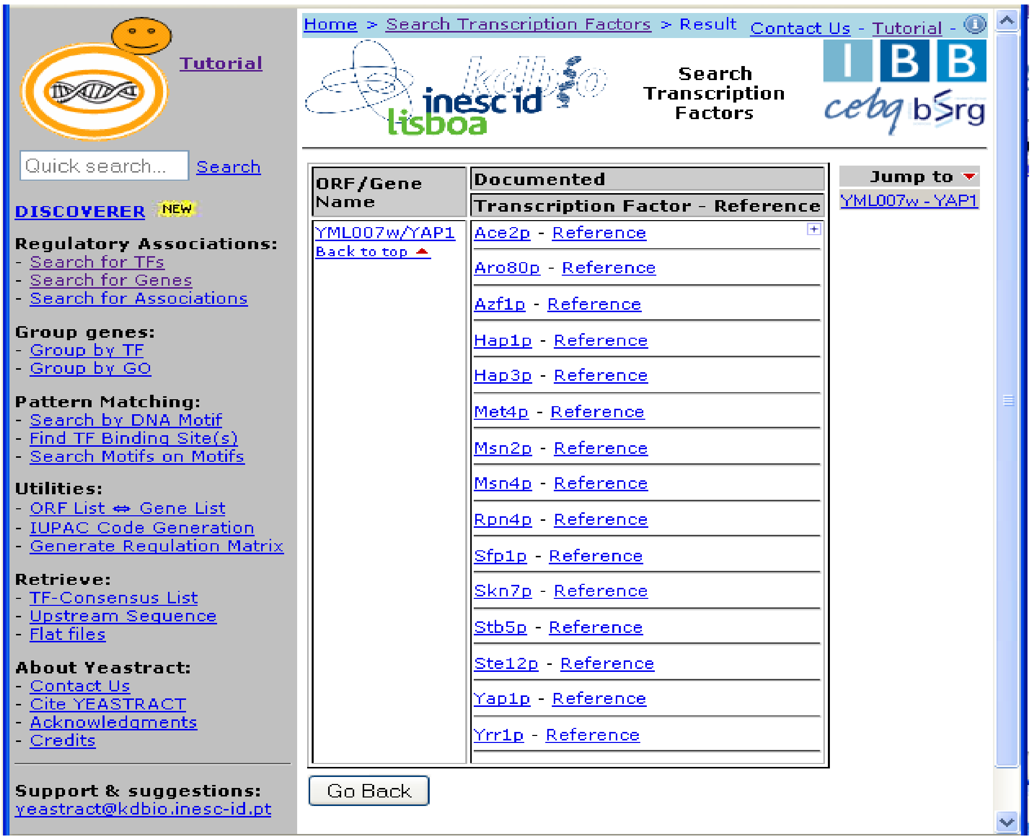
The existence of databases where biological findings are maintained in a structured and standardized manner enables a faster and efficient retrieval, exchange and analysis of data. This is also true for data concerning gene transcription regulations, and was the reason behind the development of Yeastract – the YEAst Search for Transcriptional Regulators And Consensus Tracking database.

This chapter provides a brief description of Yeastract and related databases, as well as of other databases relevant for the development of GREAT.

3.1 Yeastract

Yeastract (Figure 1) was created by INESC-ID [19] and the Biological Sciences Research Group from the *Centro de Engenharia Biológica e Química - Instituto Superior Técnico* [20].

This database contains regulatory associations between TFs and target genes in *S. cerevisiae*, manually curated from the literature. Table 2 contains information regarding the amount of data stored in Yeastract when it was created (2006) and currently (as of September 2008). It can be seen that the number of gene regulations increased almost three fold, accompanied by an increase of 28% in the number of bibliographic references.



ORF/Gene Name	Documented Transcription Factor - Reference
YML007w/YAP1 Back to top	Ace2p - Reference
	Aro80p - Reference
	Azf1p - Reference
	Hap1p - Reference
	Hap3p - Reference
	Met4p - Reference
	Msn2p - Reference
	Msn4p - Reference
	Rpn4p - Reference
	Sfp1p - Reference
	Skn7p - Reference
	Stb5p - Reference
	Ste12p - Reference
	Yap1p - Reference
	Yrr1p - Reference

Figure 1. Screenshot of Yeastract exemplifying a search query for the TFs that regulate the gene Yap1.

Table 2. Volume of data stored in Yeastract database, on 2006 and 2008.

	2006	**Sept 2008**
Regulatory associations	12,346	34,518
Bibliographic references	861	1,099
DNA-binding motifs	257	284
Unique binding motifs[1]	181	208

[1]- This count refers to binding motifs specific for no more than one TF.

Some of the information used to populate Yeastract has been obtained from external databases: data concerning genes from SGD; data about gene annotations from GO; and data on nucleotide sequences (of coding regions and promoters) from Regulatory Sequence Analysis Tools [6].

The regulations contained in Yeastract are catalogued either as documented or potential. They are documented when the regulation was identified with methods that either analyze the binding of the TF to the target gene promoter region or the changes in the target gene expression in consequence to the transcription factor suppression; and are potential when the only experimental evidence found was the transcription factor binding motif in the promoter region of the target gene [6].

Yeastract includes tools for several tasks related with the stored regulatory associations:

- Identification of complex motifs found to be over-represented in the promoter regions of co-regulated genes.
- Comparison between DNA motifs and the TF binding sites described in the literature.
- Identification of documented or potential transcription regulators of a given gene and of documented or potential genes regulated for a given TF.
- Grouping of a list of genes (for instance a set of genes with similar expression profiles) based on their regulatory associations with known transcription factors [6].

3.2 Saccharomyces Genome Database

Yeastract references the SGD for the obtention of further information about genes.

SGD is a scientific database of the molecular biology and genetics of the yeast *S. cerevisiae*, housed in the Department of Genetics at the School of Medicine, Stanford University. It contains the following data:

- Sequences of yeast genes and proteins.
- Descriptions and classifications of the biological roles, molecular functions, and subcellular localizations of genes and proteins.
- Links to literature information.
- Links to functional genomics datasets.
- Tools for analysis and comparison of sequences [21].

The database curators maintain a list of categories that describe the kind of biological information that the papers may contain, and assign one or more of these categories to each paper during the curation process. The categories refer both to specific chromosomal features and to more general information about yeast. The following is a list of examples of both category types:

- DNA/RNA Sequence Features - DNA sequence and sequence features (promoters, exons, introns, etc.), and RNA sequence features (splice sites, poly-A sites, etc.).

- Function/Process - the role played by the protein in the cell and function specifications (for example, what type of enzyme it is).
- Evolution – refer to studies that discuss *S. cerevisiae* evolution in general, as well as evolutionary studies of specific *S. cerevisiae* genes.
- Genomic expression study - includes microarray/chip/serial analysis of gene expression (SAGE) or other genome-wide techniques to assay gene expression on a genomic scale [21].

This database is used in GREAT for the obtention of data referenced in the papers: the gene names, necessary for the identification of the putative regulations; and the method used for the identification of the regulations, which is used in the feature Experimental Evidence.

3.3 Gene Ontology Database

Yeastract also references the GO database for the obtention of further information concerning functional annotations of all genes.

GO was created due to the need to describe and conceptualize biological entities in a non ambiguous manner, providing consistent functional annotations of gene products in a species-independent fashion.

The GO project developed three structured controlled vocabularies, independent of each other, to describe gene products: Molecular Function, Biological Process and Cellular Component. A given gene product executes a certain biochemical action (Molecular Function) as a part of a biological process, in a specific cell compartment [11].

GO is structured as a direct acyclic graph, which means that there exist multiple parent-child relationships between the terms that compose the ontology but that cycles cannot exist. The root term in the ontology is *all* (Figure 2), of which *molecular function*, *biological process* and *cellular component* are children. The ontology is structured in such a way that the terms nearer to the graph's root provide less specific information about the gene products annotated with them; the terms' specificity increases along a path, with the leaf terms (the last in the path) having the highest specificity. For a given term to be introduced in the ontology, it has to respect the true path rule that states that “the pathway from a child term all the way up to its top-level parent(s) must always be true” [22].

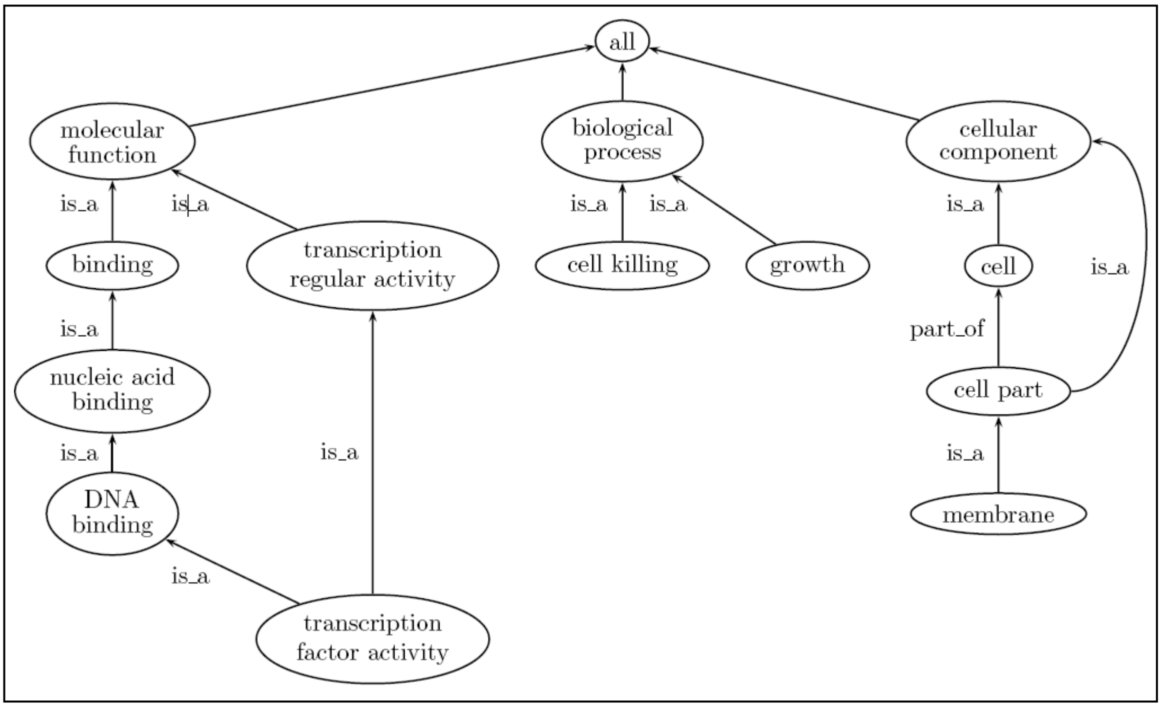


Figure 2. Exemplifying representation of the graph structure of GO. The following aspects can be seen: the root term *all* and its children *molecular function*, *biological process*, *cellular component*; two of the relationships types, *is-a* and *part-of*, as well as their directionality.

The relations between terms can be expressed in three different manners:

- *is_a* – refers to a class-subclass relationship.
- *part-of* – refers to a part-whole relationship.
- *regulates* – refers to a relationship where one process directly affects the manifestation of another process (or quality).

The terminology defined by GO, and with which the gene products are annotated, is used by GREAT for the obtention of the feature Biological Potential. This feature is based on the semantic similarity of the ontology terms shared by the gene and the transcription factor in a putative regulation.

The following section contains a description of the concept ‘semantic similarity’ and of some approaches to calculate it.

3.3.1 Semantic Similarity

Semantic similarity measures provide a means to estimate how related in meaning two concepts are. Considering a measure that provides values in the interval [0,1], if the semantic similarity between two concepts is close to ‘1’ it signifies that they are highly related, and if it is close to ‘0’ it signifies that they are distantly related.

It is possible to compare gene products with a semantic similarity measure using their ontology annotations. Several measures have been devised for use with GO since the comparison of gene products at a functional level is important for several applications, and GO is widely adopted by the scientific community.

Many of the existent semantic similarity measures are based on the notion of information content (IC). The IC of a concept is based on the probability of usage of the concept in a corpus [23]: a term that occurs less often is considered more informative than one that occurs more often. Measures based on the IC rely on the notion that the similarity between two concepts can be given by the extent to which they share information [24].

3.4 Gene Ontology Annotation Database

The Gene Ontology Annotation Database (GOA) is housed by the European Bioinformatics Institute and aims to provide high-quality GO annotations to proteins in the UniProt Knowledgebase (UniProtKB) and the International Protein Index (IPI), being also a central dataset for other major multi-species databases [25].

GOA became a member of the GO Consortium in 2001, and is responsible for the integration and release of GO annotations to the human, chicken and cow proteomes, although due to the multi-species nature of the UniProtKB it also assists in the curation of another 120,000 species [25].

In GREAT, GOA is used as the source of annotation data, and as a basis to calculate the information content of GO terms [26].

3.5 Other Resources

An extensive search was performed in order to identify databases containing the experimental methods referenced in papers in a format amenable to computation. Both generic *S. cerevisiae* databases and methodology databases were queried, and three relevant databases were found: SGD, ArrayExpress [27] and Gene Expression Omnibus [28,29]. Of these, SGD was chosen as it contains references to more methodologies and lists a higher number of papers annotated with them.

In order to obtain the genes referenced in papers, the first choices were the Entrez-PubMed and Entrez-Gene databases [29], which together contain this information. However, the SGD was later found to have a higher number of genes listed per paper than the Entrez databases.

The following sections contain more information concerning the databases introduced here.

3.5.1 Entrez Databases: PubMed, Gene and Gene Expression Omnibus

Entrez is a retrieval system developed by NCBI with the purpose of performing text-based searches in their multiple databases at a time.

Two of those databases are PubMed – a literature database containing abstracts in scientific fields as medicine and preclinical sciences - and Gene – a molecular database which information on genomes' sequences and annotations [29].

Gene Expression Omnibus is a repository for heterogeneous data sets from high-throughput gene expression and genomic hybridization experiments. It is also possible to query this repository with the Entrez system, through the GEO Profiles and GEO Datasets databases.

3.5.2 ArrayExpress

ArrayExpress, developed by EBI, is a public archive for functional genomics data obtained from array based platforms, including gene expression and chromatin immunoprecipitation experiments.

The three major goals of this repository are: to serve the scientific community as an archive for data supporting publications; to provide easy access to high-quality data in a standard format; and to facilitate the sharing of technical platforms, specifically microarray designs and experimental protocols [27].

All the databases described in this chapter are public and are either used in the implementation of GREAT or where identified as alternative data sources. Table 3 contains the links for these databases.

Table 3. Links for the databases described in Chapter 3.

Yeastract	http://www.yeastract.com/
Databases used in GREAT:	
SGD	http://www.yeastgenome.org/
GO	http://www.geneontology.org/
GOA	http://www.ebi.ac.uk/GOA/
Alternative databases to use in GREAT:	
Entrez Databases	http://www.ncbi.nlm.nih.gov/Database/
ArrayExpress	http://www.ebi.ac.uk/microarray-as/ae/

Chapter 4

Design and Implementation

GREAT is not a stand-alone tool: it is a component of a system whose purpose is to identify papers containing gene regulations so these can be manually extracted by Yeastract curators.

The first section of this chapter describes briefly the text mining component that identifies the papers whose abstracts reference one or more TFs, providing the input to GREAT. The next sections describe in detail the implementation of GREAT.

4.1 Identification of TFs in Scientific Literature

The software for this component is being developed in Python and comprises four modules responsible for the following tasks [30]:

- Obtention and storage of abstracts.
- Identification of TFs in the abstracts.
- Identification and score attribution to selected text features (used to build a statistical model).
- Classification of the abstracts as relevant or non-relevant for the purpose of gene regulations, using libbow's implementation of Support Vector Machines (SVM) [31].

All abstracts are obtained from the literature database PubMed and, for training purposes, were selected as follows:

- Positive set - abstracts of papers used to populate Yeastract database, hence known to contain one or more gene regulations.
- Negative set – since a curated set of negative instances (papers not containing gene regulations) does not exist, a set of abstracts of papers referring only the Saccharomyces genus was used as a pseudo-negative set.

The rationale behind the design of this component was that a sentence referring a TF may also refer a gene regulation in which it participates. Evidently this is not always true, but the module that identifies and scores text features was designed precisely to evaluate if the words around a TF can be interpreted as a description of a gene regulation.

Each sentence containing a TF is considered an instance, of which the text features are used to train the SVM. Sentences containing a TF and a possible description of a gene regulation are considered as positive, the remaining sentences are considered as negative.

4.2 GREAT Design

From the output file of the SVM are selected and extracted the instances identifying the PubMed Ids of the papers to be further analysed by GREAT, as well as the TFs present in the abstract of each of those papers.

The genes referenced in each paper are obtained from a public biological database and a pairwise combination of the genes with the TFs is performed, in order to identify all the potential gene regulations described in the paper.

Since not all of the identified gene-TF pairs correspond to actual regulations, a confidence score is attributed to each pair, based on the output of the following features:

- Biological Potential – if a gene is regulated by a TF, they are expected to participate in the same biological processes. This feature provides a measure of how similar the biological processes are for a given pair (continuous output).
- Physical Potential – the transcription of a gene can only be directly regulated by a TF if that TF binds to the promoter region of the gene. This feature is indicative of the existence of the TF binding motif in the promoter region of the gene, and therefore of the possibility of the existence of physical binding (binary output).
- Experimental Evidence – regulations identified with direct experimental methods are necessarily direct regulations, but the same is not true for those identified with indirect methods (which have a higher likelihood of not being true (direct) regulations). This feature is indicative of whether the method used to identify the regulation is direct or indirect (binary output).

The final confidence score is calculated with a machine learning method that combines the outputs of the previous features. The requirements for this method were two-fold: capability to produce a numeric output and simplicity so that it can be expeditiously implemented and interpreted. According to these requirements, regression trees [32] and model trees [33,34] were the first choices since they produce the desired output; their implementation requires the manipulation of few parameters; and the logical growth of a tree is based on linear divisions of the space of solutions, providing an easy-to-understand representation of the partitions made by the algorithm.

4.3 GREAT Implementation

4.3.1 Databases

Two databases were used in the development of this work: ProteInOn [26], which integrates the GO and GOA databases; and Yeastract.

Yeastract (version from September 2008) was locally installed using MySQL and was named ARN. Two tables were added to this database (Figure 3), exclusively for the implementation of GREAT: one containing gene data (common name, systematic name and a protein identifier from ProteInOn - protId) and the other containing data on the gene-TF pairs (PubMed Id, gene and TF internal identifiers, output for each one of the features described in the previous sub-section and confidence score).

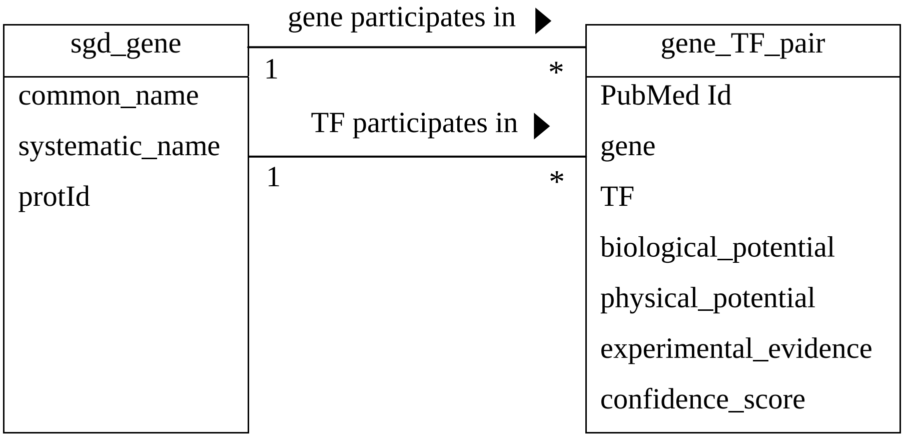


Figure 3. UML schema of the tables created for the implementation of GREAT, sgd_gene and gene_TF_pair. The attribute 'protId' stored in the table sgd_gene is an external identifier obtained from the table Prot_Info in the database ProteInOn.

The tables containing the data about the genes and the pairs are duplicated in order to accommodate training data and new data (to be classified by the trained tool).

The only difference between them is that the training table contains a field for the type of instance (positive or negative), instead of the confidence score.

4.3.2 GREAT building blocks

This evaluation tool is composed of a total of six Perl modules, five of which are common to training and new data, and one ('DefineTrainingSets') is specific for training data (Figure 4).

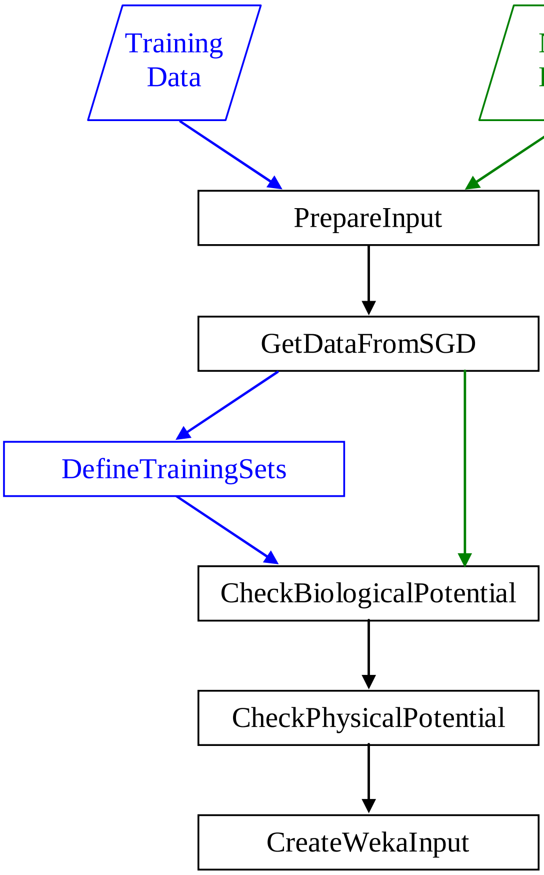


Figure 4. Fluxogram representing the workflow of GREAT. The modules used both by training and new (unclassified) data are depicted in black. Training data is depicted in blue, with the representation of its entrance into the workflow and its passage through the module 'DefineTrainingSets'. New (unclassified) data is depicted in green, with the representation of its entrance into the workflow and its direct course from the module 'GetDataFromSGD' to 'CheckBiologicalPotential'.

The first step in GREAT is the obtention of its input from the text mining tool output. While training data is obtained from the SVM output obtained with training data, new data is obtained from the SVM output obtained with unclassified abstracts. As mentioned in section 4.1, the data used to train the SVM includes sentences with TFs identified in the abstracts of papers used to populate Yeastract.

In the case of the obtention of the training data for GREAT, and after the identification of the SVM settings that produced the best results, a total of 50 runs were performed using 60% of the instances for training and 40% for testing. This resulted in an output file of the SVM containing a compilation of all instances used to test the models learned, and their final classification (either positive – sentences containing a TF and a possible description of a gene regulation – or negative). Since the instances are used more than once, only those classified as positive in every run were selected as potentially describing a gene regulation.

In the case of the obtention of new data for GREAT, only one run of the SVM is to be performed for each batch of unclassified abstracts, and the instances selected are those whose final classification is positive.

The module ‘PrepareInput’ was designed to produce the input of GREAT. For each instance selected from the output of the SVM, the module performs the extraction of the PubMed Id of the paper to which the instance belongs and of the TFs it refers.

The second step includes the obtention of the genes referenced in the papers to further analyse and the Experimental Evidence feature. This is performed by the module named ‘GetDataFromSGD’, which accesses a file downloadable by ftp from SGD [35]. Since these files are continually updated, they have to be periodically downloaded (the module receives the new file name as a parameter). The version used during the development of this work is from February 2009.

The file from SGD contains literature information that includes the following: PubMed Id, the bibliographic reference, the gene names (common and systematic), and a list of categories describing the biological information contained in the paper.

The gene names are directly stored in the local database, ARN, as well as the gene-TF pairs obtained with the pairwise combination of genes with TFs.

From this point on, the actions performed by the modules of GREAT are centered in the gene-TFs pairs.

The Experimental Evidence is obtained from the list of categories that describe the biological information. Amongst the existent categories, three refer to methods used for the identification of gene regulations: “Genomic co-immunoprecipitation study” (includes ChIP and ChIP-chip assays – direct methods), “Genomic expression study” (microarrays – indirect method), “Large-scale protein detection” (proteomics – indirect methods). The Experimental Evidence is a binary feature: gene-TF pairs from papers describing a direct method receive the value 1; pairs from papers describing an indirect method receive the value -1. The feature’s value is a missing value when no method is identified.

The module ‘DefineTrainingSets’, the only specific for training data, identifies each gene-TF pair as a positive or negative instance, for the purpose of training GREAT. Positive instances, labelled 1, correspond to pairs identified by GREAT that are present in Yeastract as documented regulations (experimentally confirmed); negative instances, labelled 0, correspond to pairs identified by GREAT but that are not present in Yeastract.

The module ‘CheckBiologicalPotential’ performs the calculation of the Biological Potential. This potential is a value in the interval [0,1] and corresponds to the semantic similarity between a TF and the gene it potentially regulates. This similarity is calculated using the GO Biological Process terms with which the TF and gene are annotated in the GOA database (both manually curated and electronic annotations are considered).

The semantic similarity measure used is IC-based and is an extension of Resnik's measure [24] for comparing genes or proteins (rather than terms). Starting with the list of all the terms annotated (directly or by inheritance) to the gene and the TF, the terms they share are identified and the term with highest IC is selected from these [36]. Therefore, the Biological Potential value is the IC of the most informative (or specific) term shared by the TF and the gene. The higher this value, the more specific is the biological process shared by the TF and gene, and the more likely it is that they are related.

The selection of this specific semantic similarity measure was based on the fact that existent comparative studies consider it as the most successful in terms of protein/protein interaction prediction and/or validation [37]

The Physical Potential is obtained in the module ‘CheckPhysicalPotential’. First it is verified if the gene-TF pair is present in a list of potential regulations from Yeastract. A regulation is potential when the only experimental evidence found was the presence of the TF binding motif in the gene promoter region. Yeastract stores the information of what genes contain which TF binding motifs in this list, but not for regulations already documented.

Secondly, if the pair is not found in the referred list, a match is performed between the gene promoter region and the TF binding motif. This match consists in a simple verification if the promoter sequence (a string of letters) contains the TF motif (a substring of letters). If the TF can physically bind the gene, the Physical Potential of the pair is 1 otherwise it is -1.

The promoter sequences of the genes and the TF binding motifs are obtained from Yeastract. For some TFs the motifs contain degenerate nucleotides (that is, symbols that represent a position in a DNA sequence that can be one of multiple nucleotides), but only those with non-degenerate nucleotides are used to perform the match.

The last module, ‘CreateWekaInput’, accesses the data stored in the local database ARN and writes a file in a specific format (arff) to be used by Weka [38]. This is a collection of machine learning algorithms for data mining tasks that includes an algorithm that can create both model and regression trees, the M5’ [34].

Confidence score calculation

As explained before, the algorithm M5’, implemented in Weka, creates model and regression trees. Weka presents this possibility as a definable parameter, among the following: minimum number of instances allowed in a leaf node and use of pruning.

In terms of applicability, both model and regression trees calculate a numeric output, with the difference between them residing in the format of that output. While regression trees store in each leaf a class value that represents the average value of instances that reach that leaf, model trees store a linear regression model that predicts the class value of the instances that reach that leaf [38].

The confidence score was only calculated for the training data. The results obtained were not suitable to decide which type of tree to use, despite the variations of the algorithm parameters tested. This question will be addressed in detail in the following chapter.

Chapter 5

Results and Discussion

This chapter describes three types of results:

- Those obtained from the training of regression and model trees using the algorithm M5’.
- Those obtained from the analysis of the data used to train the trees.
- Contributions of this work to the scientific community.

5.1 Regression and Model Trees

The data used to train and test the trees was the same, performing a 10-fold cross-validation. The variation of the minimum number of instances accepted in a leaf was tested, as was the use of pruning. When considering 4 instances in a leaf, the values presented correspond to three runs, and when considering 8 and 16 leaves, they correspond to only one run.

Table 4. Statistics of the results obtained in the implementation of the regression tree, when performing 10-fold cross-validation with the training data. Number of instances: the minimum number of instances in each leaf; RMSE: root mean squared error; CC: correlation coefficient; Leaves: number of leaves in the final tree. The values indicated correspond to the arithmetic mean of 3 runs when using 4 instances, and to a single run when using 8 and 16 instances.

Number of Instances		RMSE (%)	CC	Leaves
4 Instances	**Pruning**	45,5	0.4136	22
	No Pruning	44,0	0.4795	133
8 Instances		45, 6	0.4120	22
16 Instances		45,7	0.4053	22

The statistics considered for comparing the performances of the trees are the root mean squared error (RMSE), which is referred by Witten and Frank (2005) as a good criterion for regression, and the correlation coefficient, that measures the statistical correlation between the predicted and the actual values of the instances. It is expected of a well trained tree (or any other method) that the error be has low as possible and the coefficient correlation as closer to '1' as possible.

As can be seen in Table 4 and Table 5, the error values obtained for both methods are around 40-50% and the correlation values are below 0.5. Furthermore, neither of the statistics appears to be visibly influenced by the variations of the parameters, with the exception of pruning. When no pruning was performed, the RMSE decreased slightly and the correlation coefficient increased. However, due to the high number of leaves in the resultant tree which decrease its interpretability and due to the risk of overfitting, the absence of pruning is not desirable.

Since the obtained values of RMSE and correlation are not the expected for a good trained tree, it was not possible to choose one of the methods, regression or model trees, for the calculation of the confidence score. In order to verify if the results obtained are due to the inadequacy of the models tested or due to the data used to train them, a further analysis of the training data was performed.

Table 5. Statistics of the results obtained in the implementation of the model tree, when performing 10-fold cross-validation with the training data. Number of instances: the minimum number of instances in each leaf; RMSE: root mean squared error; CC: correlation coefficient; Leaves: number of leaves in the final tree. The values indicated correspond to the arithmetic mean of three runs when using 4 instances, and to a single run when using 8 and 16 instances.

Number of Instances		**RMSE (%)**	**CC**	**Leaves**
4 Instances	**Pruning**	45,3	0.4244	20
	No Pruning	43, 6	0.4929	133
8 Instances		42,2	0.4534	20
16 Instances		45,4	0.4188	20

5.2 Training Data Analysis

From the text mining output obtained with training data, GREAT selected 205 papers whose abstracts contain at least one TF, and that potentially describe a gene regulation. For 195 of these papers, the names of the genes referenced therein were obtained. Table 6 shows the resulting number of genes, TFs and gene-TF pairs.

The pairs identified as ‘Positive’ correspond to gene regulations existent in Yeastract and those identified as ‘Negative’ correspond to regulations identified by GREAT, but not existent in Yeastract. It is important to have in mind that some of the pairs considered ‘Negative’ can be false negatives. This is possible since the negative training set used in the text mining tool might contain abstracts that contain regulations, but whose papers were never curated by Yeastract curators. This is one aspect that can hamper the performance of GREAT, due to the existence of miss-classified training data.

Table 6. Data statistics for GREAT: number of regulation pairs (with indication of the number of positive and negative instances), total number of genes, number of TFs.

Gene-TF pairs			Genes (including TFs)	TFs
Positive	**Negative**	**Total**		
916	918	1834	635	90

Table 7 shows the characterization of the training data in terms of the three features: Biological Potential average, Physical Potential and Experimental Evidence frequencies, and missing values for each feature.

Table 7. Training data descriptors. For each set of pairs, positive and negative: Biological Potential average, Physical Potential and Experimental Evidence frequency counts; missing values percentage.

Feature		Positive Pairs	Negative Pairs	Missing values (%)
Biological Potential	Average	0.31	0.36	3.9
	Std. Deviation	0.25	0.24	
Physical Potential	Yes	287	116	36
	No	298	466	
Experimental Evidence	Direct	0	0	76
	Indirect	272	176	

Biological Potential

The Biological Potential measures the similarity between the biological processes in which the gene and TF in a given pair participate. Considering that a TF is likely to be involved in some of the biological processes in which the gene it regulates is involved, positive pairs were expected to have higher Biological Potential values than negative pairs. However, this was not true for the training data, where the negative pairs had a slightly higher average Biological Potential (Table 7).

Since the average values are relatively low (below 0.5) for both positive and negative pairs, it is possible that most genes and TFs are annotated with not very specific terms, and thus positive and negative pairs are difficult to distinguish using a semantic similarity measure. However, it is also possible that the semantic similarity measure used is not the most adequate for this data, and that further information could be extracted with an alternative measure.

The possible existence of false negatives in the training set may also be a factor behind the similar Biological Potential values obtained for the positive and negative pairs, since the false negatives are expected to have Biological Potential values as high as those observed for the positives.

For this potential the missing values are negligible (3.9%).

Physical Potential

Since the Physical Potential indicates whether a given TF can physically bind to a gene (a prerequisite of a direct regulation) it is expected that the majority of positive pairs have a positive potential and that the majority of the negative have a negative potential. However, as can be seen in Table 7, for the positive pairs there is a similar number of cases with positive potential and of cases with negative potential. There are two possible explanations for the existence of so many positive pairs with no physical potential: either the gene is indirectly regulated by the TF, and so the physical binding does not occur; or the method used to evaluate the existence of the transcription factor binding motif in the gene does not work as desired. This last explanation is based on the fact that binding motifs containing degenerate nucleotides have not been considered in the match between binding motif and promoter region. As such, the identification of the true number of pairs with positive potential may be diminished.

In the case of the negative set, only 25% of the pairs have a positive potential. This percentage might be due to the existence of false negative pairs, or to the fact that the existence of the binding motif in the promoter of the gene does not determine the validity of the regulation itself.

For this potential, the missing values are considerable (36%). They correspond to pairs for which the promoter sequence of the gene, the TF binding motif, or both did not exist in Yeastract.

Experimental Evidence

This feature was the most problematic to implement due to the difficulty in finding databases that referenced the methods present in papers in a manner suitable for computational extraction.

The analysis of the training data allowed the identification of two major problems associated with the Experimental Evidence:

- There are no gene-TF pairs identified with direct methods, only with indirect ones.
- The missing values amount to 76% of the pairs.

The inexistence of direct methods represents an important drawback for a training dataset, and implies that this feature has no ability to separate negative from positive pairs.

Given the limitations encountered for this feature, it is not possible to draw any hypothesis about the influence of the possible existence of false negative pairs.

The missing values correspond to pairs for which it was not possible to obtain the method from the SGD database.

Table 8. Example cases of false negative pairs. For both pairs the values of the features are indicative of a regulation.

Gene	TF	Physical Potential	Biological Potential	Experimental Evidence	PubMed Id
Rtg3	Rtg1	1	0.881557	Missing value	17351075
Msn4	Msn2	1	0.908332	Missing value	10409737

In Table 8 are presented two examples of gene-TF pairs classified as negative pairs, but whose features indicate that they might correspond to a true regulation. From the names of the papers to which both pairs correspond, it is expected that the papers contain descriptions of gene regulations:

- PubMed Id 17351075 – "Multiple basic helix-loop-helix proteins regulate expression of the ENO1 gene of Saccharomyces cerevisiae".

- PubMed Id 10409737 – “Osmotic stress-induced gene expression in Saccharomyces cerevisiae requires Msn1p and the novel nuclear factor Hot1p”.

More examples of cases similar to these can be seen in the Attachment.

5.3 Contributions of this work to the scientific community

The development of the present work resulted in the implementation of a first prototype of GREAT. The following functionalities of this prototype are fully implemented:

- Module ‘GetDataFromSGD’ - Browse of a SGD file containing literature information (specific SGD file format); identification of contents using the PubMed Id and extraction of the gene names and experimental evidence labels.
- Module ‘CheckBiologicalPotential’ – Calculation of the semantic similarity between two gene products. This module is optimized to use with ProteInOn, and depends on data previously obtained in the module ‘GetDataFromSGD’.
- Module ‘CheckPhysicalPotential’ – Evaluation of the presence of a TF binding motif in the promoter region of a gene. This module is optimized to use with Yeastract, and depends on data previously obtained in the module ‘GetDataFromSGD’.

Although these modules were implemented sequentially and are better suited to be used as described in section 4.3, only minor changes would be required in order to embed them into another system.

The features obtained by GREAT for the positive training regulations are stored in a relational database and can be used for other applications, since they correspond to manually curated regulations from Yeastract.

The analysis of the training data allowed the identification of implementation aspects of this first prototype that are not working as desired (since the results obtained are unsatisfactory from the biological point of view), and of possible solutions to this problems.

Chapter 6

Conclusions and Future Work

Yeastract was created to provide the *S. cerevisiae* scientific community with a public database of transcription regulatory associations. The data with which the database is populated is obtained by human curators that identify the papers describing the regulations and then read them to extract that information.

With the astounding rate at which papers are being published, it is impossible to depend solely in a manual curation process to keep Yeastract up to date. This has prompted the development of automatic tools to help in this process, wherein GREAT is inserted.

GREAT is part of a system whose purpose consists in the identification of papers describing *S. cerevisiae* gene regulations, solely using public datasources: the papers' abstracts and public biological databases. The first component of this system uses text mining to identify the papers sought through the identification of transcription factors in theirs abstracts and the evaluation of the words around the transcription factor, verifying if they are likely to describe a regulation. The second component is GREAT, which receives the list of papers selected by the first component, and that, after obtaining the genes referred in the paper from Saccharomyces Genome Database, evaluates if any of the gene-TF pairs found corresponds to a true regulation. The evaluation results in the assignment of a confidence score to each pair, and those papers that contain putative regulations with high scores will be manually curated to extract the regulations and store them in Yeastract. The score attribution is based on three features: Biological Potential, Physical Potential and Experimental Evidence. These depend on data obtained from public biological databases – respectively Gene Ontology, Yeastract and Saccharomyces Genome Database. The output of all features is combined with a machine learning method, either a regression or a model tree, which calculates the confidence score.

After the obtention of a training dataset, with positive gene-TF pairs corresponding to regulations identified in Yeastract and negative pairs corresponding to relations not identified in Yeastract, that data was used to train both types of trees, regression and model. Due to the low quality of the results obtained, it is not yet clear which of the trees, if any, might be best suited for this type of data. These results, obtained for both the methods, and presenting no visible differences upon parameters variations, prompted the need to analyse the data in detail.

It is important to bear in mind that the pairs considered as negative might include false negatives, since the negative dataset used to train the text mining tool might include abstracts containing regulations not extract by Yeastract curators.

The analysis performed over the training data revealed the following issues concerning the three features, respectively Biological Potential, Physical Potential and Experimental Evidence:

- The similarity between the biological processes in which the gene and the TF participate is, in average, relatively low (below 0.5, in a scale between 0 and 1). Since this happens for positive and negative pairs alike, the utilization of a semantic measure to help in their separation is difficult. This might be due to the fact that the biological entities are annotated with not very specific terms or that the semantic similarity measure used is not the best suited.

  Since the average values of positive and negative pairs is very similar, it is possible that the existence of false negatives might be responsible for higher average values for negative pairs.

  This means that very little information can be retrieved from these values, hampering the capacity to infer if the gene and the transcription factor really participate in the same biological process and, consequently, if the regulation of the gene by that transcription factor is biologically possible.

- About 50% of the pairs identified in Yeastract as documented regulations have a negative Physical Potential. The physical binding of the TF to the gene is imperative in direct regulations, and this information is used to assist in the validation of a gene-transcription factor pair as a true regulation. It is possible that the match process is not identifying all possible pairs with a positive potential as a consequence of the use of only binding motifs that do not contain degenerate bases.

- The number of gene-transcription factor pairs for which no experimental methods was identified is as high as 76%. In addition, for the remainder pairs the method identified is indirect. This renders this feature useless for the pairs' evaluation.

In order to deal with the problems exposed by the data analysis, there are some aspects in the implementation of GREAT that might be differently approached:

- The papers that contain pairs identified by GREAT and considered negative regulations should be analysed to verify if the pairs correspond to true negatives or false negatives.
- The genes referenced in the papers may be obtained from the Saccharomyces Genome Database and also from the Entrez-Pubmed and Entrez-Gene databases (aiming not only at a higher number of genes per paper but also at a higher number of papers annotated).
- The Biological Potential can be calculated using a different semantic similarity measure, one that, for instance, considers more than just one shared term, providing a more global perspective of the similarity between the genes and the TFs.
- The Physical Potential can be evaluated considering transcription factors' binding motifs containing degenerate nucleotides and, when necessary, databases other than Yeastract may be used to obtain the promoter sequences and the transcription factors binding motifs.
- The Experimental Evidence can be obtained from the Saccharomyces Genome Database and also from Gene Expression Omnibus and ArrayExpress. Although the first contains more methods and a higher number of papers annotated with the methods, the combined use of all three databases may decrease the missing values.

Not only the currently existent features of GREAT may be manipulated, but also new features may be considered:

- It is possible to verify in the abstract if a gene name is in the vicinity of a transcription factor (using the text features with which the text mining tool evaluates if a sentence describes a regulation).
- Knowing that a gene in a pair participates in the same biological process as another gene that is known to be regulated by the transcription factor

considered in the pair, it is conceivable that the pair may correspond to a regulation.

The contributions of the present work can be resumed as follows: successful implementation of a first prototype of GREAT, which incorporates fully implemented functionalities related with the obtention and utilization of data from external sources; the data compiled for the positive regulations is ready to use (being based on manually curated data) and is stored in a relational database; the analysis of the data used to train the prototype provided relevant information on the limitations of the system and possible solutions for these limitations.

Although the results obtained with the regression and model trees were not satisfactory from the biological point of view, the conclusions drawn about the training data analysis suggest that the features considered in the initial design are adequate to solve the problem. However, their implementation was ultimately hindered by lack of information: the semantic similarity measure used to obtain the Biological Potential does not extract enough information from the data; the binding motifs using degenerate bases were not used to obtain the Physical Potential; there are not databases containing enough information on the methods referred in the papers.

From the aspects proposed to improve GREAT, those related with the Experimental Evidence present the greatest challenge. This is due to the fact that, despite the extensive search performed, the databases containing this information have it for few papers. However, it is possible that the introduction of new features, such as those proposed above will render the use of the Experimental Evidence unnecessary.

Bibliography

[1] Haring M., Offermann S., Danker T., Horst I., Peterhansel C. and Stam M.: Chromatin immunoprecipitation: optimization, quantitative analysis and data normalization. Plant Methods 3, 2007. doi: http://dx.doi.org/10.1186/1746-4811-3-11 URL: http://www.plantmethods.com/content/3/1/11.

[2] http://www.promega.com/guides/protein.interactions_guide/chap6.pdf

[3] http://www.chiponchip.org/

[4] Sheils O., Finn S. and O'Leary J.: Nucleic acid microarrays: an overview. Current Diagnostic Pathology 9, 155-158, 2003. doi:10.1016/S0968-6053(02)00095-9

[5] Tyers M. and Mann M.: From genomics to proteomics. Nature 422, 193-197, 2003. doi: http://dx.doi.org/10.1038/nature01510

[6] Teixeira M.C., Monteiro P., Jain P., Tenreiro S., Fernandes A.R., Mira N.P., Alenquer M., Freitas A.T., Oliveira A.L. and Sá-Correia I.: The YEASTRACT Database: a Tool for the Analysis of Transcription Regulatory Associations in *Saccharomyces cerevisiae*. Nucl. Acids Res. 34, 2006. doi: http://dx.doi.org/10.1093/ nar/gkj013

[7] Krallinger M., Valencia A. and Hirschman L.: Linking genes to literature: text mining, information extraction and retrieval applications for biology. Genome Biology 9 (Suppl2), 2008. doi: http://dx.doi.org/10.1186/gb-2008-9-S2-S8 URL: http:// genomebiology.com/2008/9/S2/S8

[8] Rebholz-Schuhmann D., Kirsch H. and Couto F.: Facts from Text — Is Text Mining Ready to Deliver? PLoS Biol 3(2), 2005: e65. doi: http://dx.doi.org/10.1371/journal. pbio.0030065

[9] WordNet: An Electronic Lexical Database, Edited by Christiane Fellbau. The MIT Press, 1998

[10] http://wordnet.princeton.edu/wordnet/documentation/

[11] Ashburner, M. et al: Gene Ontology: Tool for the Unification of Biology. Nature Genetics 25, 2000.

[12] Couto F. and Silva M.: Mining the bioliterature: towards automatic annotation of genes and proteins”, Advanced Data Mining Technologies in Bioinformatics, Idea Group Inc, 2006.

[13] Saric J., Jensen L.J. and Rojas I.: Large-scale extraction of gene regulation for model organism in ontological context. In Silico Biology 5, 2005.

[14] Hahn U., Tomanek K., Buyko E., Kim J.-j and Rebholz-Schuhmann D.: How feasible and robust is the automatic extraction of gene regulation events? A cross-method evaluation under lab and real-life conditions. Proceedings of the Workshp on BioNLP, pages 37-45, 2009.

[15] Gama-Castro S. et al - RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. Nucleic Acids Res. 36 (Database issue), D120-4, 2008. doi: http://dx.doi.org/10.1093/nar/gkm994

[16] http://www-mtl.mit.edu/Courses/6.050/2003/notes/chapter10.

[17] Abney S.: Partial parsing via finite-state cascades. Proceedings of the ESSLLI ’96 Robust Parsing Workshop, Prague, Czech Republic, 1996.

[18] TIGER Search visualization tool: http://www.ims.uni-stuttgart.de/projekte/TIGER/ TIGERSearch/index.shtml

[19] http://www.inesc-id.pt/

[20] http://groups.ist.utl.pt/bsrg/

[21] Cherry J.M., Adler C., Ball C., Chervitz S.A., Dwight S.S., Hester E.T., Jia Y., Juvik G., Roe T., Schroeder M., Weng S. and Botstein D.: SGD: Saccharomyces Genome Database. Nucleic Acids Res. 26(1), 1998.

[22] Pesquita C.: Improving semantic similarity for proteins based on the Gene Ontology Master Thesis, University of Lisbon, Faculty of Sciences, 2007.

[23] Resnik P.: Using information content to evaluate semantic similarity in a taxonomy. Proceedings of the 14th International Joint Conference on Articial Intelligence, 1995.

[24] Resnik P.: Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. Articial Intelligence Research 11, 1999.

[25] Barrel D., Dimmer E., Huntley R.P., Binns D, O’Donovan C. and Apweiler R.: The GOA Database in 2009 - An Integrated Gene Ontology Annotation Resource. Nucl. Acids Res. 37, 2009. doi: http://dx.doi.org/10.1093/nar/gkn803

[26] Faria D., Pesquita C., Couto F. and Falcão A.: ProteInOn: A web tool for protein semantic similarity - Technical Report. TR 07-6. DI FCUL, 2007.

[27] Parkinson, H. *et al.*: ArrayExpress update - from an archive of functional genomics experiments to the atlas of gene expression. Nucl. Acids Res. 37, D868-D872, 2009. doi: http://dx.doi.org/10.1093/nar/gkn889

[28] Barrett T. and Edgar R.: Gene Expression Omnibus (GEO): Microarray data storage, submission, retrieval, and analysis. Methods Enzymol. 411, 2006. doi: http://dx.doi.org/
10.1016/S0076-6879(06)11019-8

[29] Sayers E.W. *et al*: Database Resources of the National Center For Biotechnology Information. Nucl. Acids Res. 37, 2009. doi: http://dx.doi.org/10.1093/nar/gkl1031

[30] Couto F., Grego T., Pesquita C., Bastos H., Torres R., Sanchez P., Pascual L. and Blaschke C.: Identifying bioentity recognition errors of rule-based text-mining systems. IEEE Third International Conference on Digital Information Management (ICDIM) 2008.

[31] McCallum A.K.: Bow: A toolkit for statistical language modelling, text retrieval, classification and clustering, 1996. URL: http://www.cs.cmu.edu/~mccallum/bow

[32] Breiman L., Friedman J., Stone C.J. and Olshen R.A.: Classification and regression trees. Chapman & Hall, USA, 1984

[33] Quinlan R.J.: Learning with continuous classes. Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, Singapore, 1992.

[34] Wang Y., Witten I.H.: Induction of model trees for predicting continuous classes. Poster papers of the 9th European Conference on Machine Learning, 1997.

[35] http://downloads.yeastgenome.org/literature_curation

[36] Cho Y.R., Hwang W., Ramanathan M. and Zhang A.: Semantic integration to identify overlapping functional modules in protein interaction networks. BMC Bioinformatics 8:265, 2007. doi: http://dx.doi.org/10.1186/1471-2105-8-265 URL: http://www.biomedcentral.com/1471-2105/8/265

[37] Pesquita C., Faria D., Falcão A., Lord P. and Couto F.: Semantic Similarity in Biomedical Ontologies. PLoS Computational Biology (in press).

[38] Witten I.H. and Frank E.: Data Mining: practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

Attachment

Hypothetical false negative pairs

Query with which these examples were obtained:

```
SELECT g.common_name, t.common_name, p.biological_potential,
p.physical_potential, p.experimental_evidence, p.pubmed_id
FROM sgd_gene_training g, sgd_gene_training t, gene_trans_factor_pair_training p
WHERE g.gene_id=p.gene
AND t.gene_id=p.trans_factor AND doc_regulation = 0 AND physical_potential = 1
AND biological_potential > 0.5;
```

Gene	TF	Biological Potential	Physical Potential	Experimental Evidence	PubMed Id
SKN7	Yap1	0.8478	1	0	10930459
RTG1	Rtg3	0.881557	1	0	17351075
MSN2	Msn4	0.908332	1	0	10411744
PBS2	MSN2	0.559752	1	0	14699125
HOG1	MSN2	0.559752	1	0	14699125
TYE7	GCR1	0.753024	1	0	15789351
MSN2	Msn4	0.908332	1	0	10409737
MSN1	Msn4	0.559752	1	0	10409737
HOG1	Msn4	0.559752	1	0	10409737
MSN1	MSN2	0.559752	1	0	10409737
HOG1	MSN2	0.559752	1	0	10409737
MSN2	Msn4	0.908332	1	0	10407268
Rtg3	RTG1	0.881557	1	0	10848632
RTG2	RTG1	0.881557	1	0	10848632
SKN7	Msn4	0.559752	1	0	11821410
MSN2	Msn4	0.908332	1	0	11821410
RAS1	Msn4	0.721201	1	0	11821410
RAS2	Msn4	0.721201	1	0	11821410
SKN7	MSN2	0.559752	1	0	11821410
RAS1	MSN2	0.721201	1	0	11821410
RAS2	MSN2	0.721201	1	0	11821410
SOK2	NRG1	0.725206	1	0	15466424
PHD1	NRG1	0.725206	1	0	15466424
KSS1	NRG1	0.744205	1	0	15466424
TEC1	NRG1	0.744205	1	0	15466424
RAS2	NRG1	0.725206	1	0	15466424
SKN7	Yap1	0.8478	1	0	16862604
SNF2	Gcn4	0.630144	1	0	12665580
SKN7	Yap1	0.8478	1	0	16313629
MSN2	Msn4	0.908332	1	-1	11102521
MSN2	Msn4	0.908332	1	0	10641036
RTG1	Rtg3	0.881557	1	0	12393187
RTG1	Rtg3	0.881557	1	0	10509019
RTG2	Rtg3	0.881557	1	0	10509019
Rtg3	RTG1	0.881557	1	0	10509019
RTG2	RTG1	0.881557	1	0	10509019
MSN2	Msn4	0.908332	1	0	9756934
RAS2	Msn4	0.721201	1	0	9756934
RAS2	MSN2	0.721201	1	0	9756934
MSN2	Msn4	0.908332	1	0	14685262
MSN2	Msn4	0.908332	1	0	15922872
MSN2	Msn4	0.908332	1	-1	10722658
HOG1	Msn4	0.559752	1	-1	10722658
HOG1	MSN2	0.559752	1	-1	10722658
MSN2	Msn4	0.908332	1	0	9649426
SKN7	Yap1	0.8478	1	0	12614847
HOG1	SKN7	0.559752	1	0	12614847
SKN7	Msn4	0.559752	1	0	12614847
MSN2	Msn4	0.908332	1	0	12614847
RAS1	Msn4	0.721201	1	0	12614847
RAS2	Msn4	0.721201	1	0	12614847
HOG1	Msn4	0.559752	1	0	12614847
SKN7	MSN2	0.559752	1	0	12614847
RAS1	MSN2	0.721201	1	0	12614847
RAS2	MSN2	0.721201	1	0	12614847
HOG1	MSN2	0.559752	1	0	12614847
HOG1	SKN7	0.559752	1	0	9843501
MSN2	Msn4	0.908332	1	0	10048026
MSN2	Msn4	0.908332	1	0	11827753
MSN2	Msn4	0.908332	1	0	11260469