Mining cis-Regulatory Transcription Networks from Literature

Florian Leitner[1*], Martin Krallinger[1], Sushil Tripathi[2],
Martin Kuiper[3], Astrid Lægreid[2], and Alfonso Valencia[1*]

[1]Structural Computational Biology Group
Spanish National Cancer Research Center, Madrid
* {fleitner,avalencia}@cnio.es

[2]Department of Cancer Research and Molecular Medicine
Norwegian University of Science and Technology, Trondheim

[3]NTNU Semantic Systems Biology Group
Norwegian University of Science and Technology, Trondheim

While transcription regulation is a key biological process, to date no public reference repository for these interactions exists. The IMEx consortium unites several protein interaction DBs, but there is no similar public repository available for Transcription Regulation Events (TREs; DNA-binding of a transcription factor to a target gene's promoter element that leads to the (up- or down-) regulation of the target). We now are building such a repository, based on TRE descriptions in the scientific literature, integrating text mining and manual curation. Annotating these interactions is a complex issue, as TREs are a subset of gene regulation events (GREs; direct or indirect regulator-target gene interactions that may involve a transcription event) that current text mining methods are extracting and go beyond the associations of transcription factors to DNA sequences derived by ChIP-seq. Detecting transcription factor DNA-binding events with their regulatory effect on target genes is a more complex issue than either text mining or ChIP-seq are currently solving. Additionally, most current text mining systems do not provide "normalized" (sequence DB-mapped) interactors, while curators spend a major part of their time resolving these identifiers.

We will present our ongoing work on an open-source, UIMA-based extraction system and the parallel effort to build a Gold Standard corpus of functional TREs consisting of directed, normalized transcription factor-target gene interactions. The key aspects of the system are: (1) its focus on transcription regulation events, (2) the integration of biologically relevant context, particularly experimental conditions and host tissue, (3) an organism-independent DB ID mapping of the participants, (4) a generic syntax expression language that enables syntactic pattern mining in UIMA, (5) an evaluation based on a hand-curated Gold Standard, (6) and the design of our curation standards for TREs to assemble this Gold Standard.

1. Introduction to Transcription Regulation Events

A significant proportion of a cell's regulatory capabilities are directed towards its RNA expression landscape (Djebali et al., 2012), with cis-regulatory transcription events at the core of the intracellular gene regulatory network (Levine and Tjian, 2003). In eukaryotes, and particularly the metazoa (multicellular eukaryotes), transcription initiation requires the binding of transcription factors (TF) to particular sequence motifs at proximal regions several hundred base pairs upstream of the target gene's (TG) transcription start site (TSS), and distal promoter regions that are further away (Lenhard et al., 2012). These events trigger the activation of RNA Polymerase II (RNA Pol II), which binds via the Pre-Initiation Complex (PIC) at the transcription start site (TSS) of protein coding and miRNA coding genes. RNA Pol II activation is mediated through a chain of protein interactions between a specific, DNA-binding TF and additional co-factors, and ultimately initiates mRNA synthesis for a particular target gene (TG) (Figure 1). In general, cellular signals are routed through these "network switches" that determine gene expression levels and thus cellular state. Given the relative sparsity (James et al., 2010) of these transcription regulation events (TREs) (Djebali et al., 2012) if compared to all possible genes a TF could pair

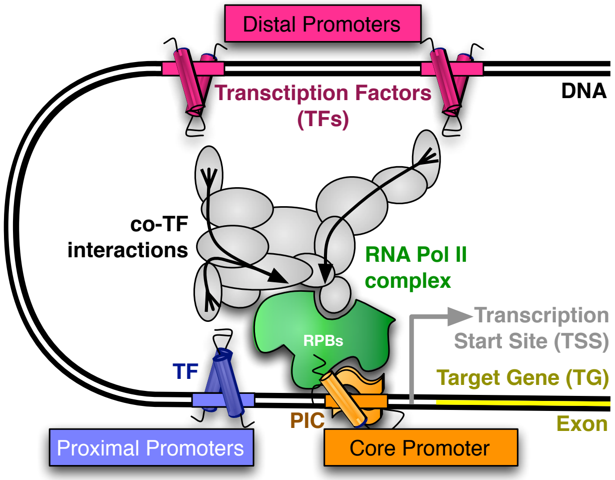

Figure 1: **A transcription regulation event (TRE).** The binding of a transcription factor (TF) to DNA in regulatory regions of the target gene (TG) leads to the regulation of RNA Pol II activity that binds via the PIC at the TSS, either directly or mediated via co-factors. A TF can be an activator or repressor (not shown) and bind distal or proximal. Downstream of the TSS, the first exon of the TG is shown.

with, a comprehensive map of the mammalian – and particularly, human – routing network is not (yet) available (Gerstein et al., 2012). However, given its central role in both normal and pathological processes, determining the transcription factor-target gene relations of gene regulatory networks is of general interest to a broad audience within the biological, medical and pharmacological sciences.

We will refer to the chain of events, from the TF binding to specific DNA regulatory regions at the TG to the particular outcome of up- or down-regulating the TG, as (direct, functional) TREs. This should be contrasted to (indirect, generic) gene regulation events (GREs), where any gene or protein can influence the expression of another (target) gene. In these cases, the actual chain of events leading to that outcome might involve several interactions via intra- and even inter-cellular signaling pathways (e.g. gene regulation in a given cell triggered by a ligand secreted from a different cell). In other words, while TREs are a subset of all GREs, the latter do not specifically describe the event of a TF directly activating or suppressing a TG, but may refer to more complex processes involving additional mediators, including intermediate transcription events. For example, the sentence "*Furthermore, p50 binds and activates the CEBPA gene in myeloid cells.*" allows an expert to infer a TRE because of the keyword *binds* and given the external knowledge that p50 is a TF. Our final goal is to extract the TF, the TG, the directionality, and cell type (p50, CEBPA, up-regulation, and myeloid cells, respectively) from this phrase. On the other hand, in "*Thus, C/EBPα and p50 reciprocally regulate each other's expression, establishing a positive feedback relationship.*", the two regulatory events (C/EBPα up-regulating p50 and p50 up-regulating C/EBPα) are clearly GREs, but the available context is insufficient to infer a TRE[1].

Another example is the report of the activation of (the GLI transcription factor family member) Gli1, mediated via aPKC-ι/λ phosphorylation, that in turn gets regulated by protein smoothened (SMO), a hedgehog (HH) receptor. This chain of events results in the transcriptional activation of the aPKC-ι/λ gene (Prkci) in basal cell carcinomas (BCCs) (Atwood et al., 2013): In the paper, with sentences such as "Prkci is a HH target gene that forms a positive feedback loop with GLI and exists at increased levels in BCCs.", the aPKC-ι/λ gene (Prkci) as well as Gli1 are presented several times as the "target genes" of hedgehog (HH). However, while the TRE is the transcriptional activation of Prkci (the TG) by Gli1 (the TF), HH and SMO are extra- and intracellular mediators of gene regulation events (GREs) with Gli1 and Prkci. Furthermore, it is probably impossible to deduce the TRE from the abstract alone, providing a cause to mine article bodies, too. Finally, the title "GLI activation by aPKC ι/λ regulates the growth of basal cell carcinomas" indicates a (reverse, protein-interaction-based) GRE, not the correct TRE.

2. Experimental TRE Detection

Over the last few years, high throughput methods have been used to trace gene regulatory networks, in particular chromatin immunoprecipitation sequencing (ChIP-seq) (Park, 2009) which detects the association of specific, genomic DNA sequences with an immunoprecipitated protein. However, determination of TREs still relies on a combination of several experimental approaches other than ChIP, many of which are mainly used in a small-scale, "low-throughput" manner, like electrophoretic mobility shift and reporter gene assays. Observations from these assays are warranted because ChIP-seq yields signals also in cases where the assayed TF binds TG regulatory regions indirectly, via another protein and thus does not ensure direct TF-DNA-binding. Furthermore, the actual transcription regulation event and its directionality (activation vs. repression) cannot be determined by ChIP-based methods (Valouev et al., 2008; Lickwar et al., 2012). To overcome such limitations, TF-sequence associations from ChIP-seq data are often combined with complementary gene expression data to predict the TGs (Sandmann et al., 2006; Huang et al., 2013). The fact that experimental data documenting functional TREs are mainly derived from small scale perturbation-type experiments performed over the last few decades implies that a wealth of TRE knowledge remains available only within the confines of a large body of published literature (A PubMed query for "*Transcription Factors[MeSH Terms] AND Regulatory Sequences, Nucleic Acid[MeSH Terms]*" on April 8, 2013, produced exactly 57,000 hits, and, for example, RegulonDB v8.0 (a yeast TRE DB) alone records 4667 papers.) Therefore, extracting these transcription events with text mining methods could provide a large number of direct, functional interactions that are not recorded anywhere else.

3. Methods for Extracting Gene Regulation Events

Despite a large number of database bio-curation efforts that are cataloging information relating to TFs and TREs[2], a comprehensive collection based on purely manual approaches will likely remain elusive; Manual curation efforts do not scale with the exponential growth of available literature (Baumgartner et al., 2007), an effect that has been observed particularly for protein interaction data/curation (Ceol et al., 2008; Leitner et al., 2010a). Therefore, a possible option to bridge the widening gap between structured TRE-related repositories and the existing literature is text-mining facilitated extraction, similar to ongoing efforts for protein interactions (Krallinger et al., 2012). The earliest GRE extraction systems date back as far as 2004 (Saric et al., 2004;

[1] both sentences taken from PubMed abstract 21346255

[2] The NAR Database issue 2013 list 76 resources for TFs and their regulator sites at http://www.oxfordjournals.org/nar/database/subcat/1/4

System (first author, year)	Pattern-based	Parser-based	Machine Learning-based	Gene Regulation Relations	Additional Relations	Target Gene Detection	Gene Normalization	Full-text IE	1 Model Organism Only
Saric et al. (2004)	■			■		■			■
Pan et al. (2004)			■		■				
Saric et al. (2006)	■			■	■	■			
Rodriguez-Penagos et al. (2007)	■			■		■	■	■	■
Fundel et al. (2007)		■		■	■	■			
Aerts et al. (2008)	■					■	■	■	
Yang et al. (2008)			■		■				
Manine et al. (2009)		■		■	■	■			■
Krallinger et al. (2009)			■	■	■	■	■	■	■
Wang et al. (2011)	■		■	■		■			
Klinger et al. (2011)			■	*	■	*		■	
Riedel et al. (2011)		■		*	■	*		■	
Roy et al. (2011)			■		■				■

Table 1: **A classification of text mining systems for gene and/or transcription regulation events.** Columns: Pattern-based: system uses (linguistic) patterns to detect relationships; Parser-based: uses dependency parsing to detect relationships; Machine Learning-based: uses statistical models to filter and/or rank (potential) interaction pairs; Gene Regulation Relations: explicitly extracts regulator-target gene relationships (*: limited; see text); Additional Relations: in addition, detects several (other) kinds of biologically relevant relationships; Target Gene: detects the target (gene) of the GRE; Gene Normalization: maps the detected regulator/TG to a database identifier; Model Organism Only: only extracts GREs for: *S. cerevisiae* (Saric, 2004 and R.-P., 2007), *B. subtilis* (Manine, 2009), *A. thaliana* (Krallinger, 2009), *M. musculus* (Roy, 2011).

Pan et al., 2004), and the former provided the basis of the STRING-IE system (Saric et al., 2006). (Hahn et al., 2009) compiled an excellent review and performance comparison of GRE extraction systems. By now, a number of text-mining systems have been published on the general topic of gene regulation event extraction; Table 1 shows several systems that have been created to detect such regulatory relationships, although not all systems shown can detect regulator-target relationships. It is important to notice that none of these systems has an explicit focus on functional TF-DNA-binding events (i.e., TREs as described in Figure 1).

With one exception, all relationship extraction systems will produce interactions and are evaluated against data that stem from descriptions of the indirect, more generic gene regulation events (e.g., pathway regulators or even extracellular peptides that trigger gene expression): Regarding TRE-based text mining systems, (Rodriguez-Penagos et al., 2007) did evaluate their text mining system against RegulonDB (yeast) data, which is comprised only of direct TF-TG interactions, although it should be noted that the regulome of yeast is less complex than that of metazoa. (Yang et al., 2008) extracted transcription factors contexts (associated GO and MeSH terms), but did not detect target genes, and the system does not normalize the TFs. (Aerts et al., 2008) applied a very successful gene normalization strategy, improving the mapping by BLASTing for sequence snippets found in the text in addition to mentions of gene symbols and names. Another noteworthy issue is that only one system (Krallinger et al., 2009) covers cell or tissue specificity for the mined interactions, which seems particularly important given its relevance in gene regulation. The work of (Wang et al., 2011) seems most related given the title, however, their evaluation was limited to only one TF family (HIF-1) in one organism (M. musculus), the patterns do not distinguish GREs from TREs, and no gene normalization is provided. Finally, (Klinger et al., 2011) and (Riedel et al., 2011) are cited as two example systems related to the BioNLP Shared Tasks (Kim et al., 2012). In this community challenge, within the "GENIA task 2", gene expression events that are combined with (positive and negative) regulatory relations could be used to deduce regulator-target entity relationships. However, the BioNLP Shared Tasks does neither implement an explicit evaluation of the pairing between regulators and targets or the mapping of these entities to their databases (known as "Gene Normalization", as for example required in the BioCreative protein interaction pair tasks (Krallinger et al., 2008; Leitner et al., 2010b)). In summary, only the systems of Rodriguez-Penagos and Krallinger come close to the objective of extracting direct, functional TRE events triggered by TF-DNA binding that lead to regulation of target gene expression reported in the form of normalized (DB mapped) TF-TG relationships, but their approaches are limited to the yeast and *A. thaliana* model organisms, respectively.

4. The `txtfnnl` Pipeline

To quote Saric (2004):

> It is often not known whether the regulation takes place at the level of gene transcription or translation or by an indirect mechanism. For this reason, and for simplicity, we decided against trying to extract how the regulation of expression takes place.

On the other hand, high-throughput ChIP-Seq is able to detect TF-DNA association events, but direct binding cannot be ascertained and it has difficulties identifying the functional transcription regulatory aspects (target gene, directionality) of a TRE. Given the complementary nature of the text mining and ChIP-seq approaches, we are creating a framework, the `txtfnnl` pipeline[3] (Figure 2), to explicitly extract direct, cis-regulatory transcriptional regulation events using text mining that could be combined with ChIP-seq data to produce an optimal set of functional TRE relations. Our target is to extract normalized (i.e., sequence database-mapped) TREs from full-text, based on patterns identified by biologists as potentially describing cis-regulatory transcription events (Table 2, top, and next section).

For the "Natural Language Processing" step in Figure 2, the OpenNLP[4] MaxEnt tagger (sentence segmentation), the GENIA Tagger (Tsuruoka et al., 2005) (part-of-speech tagging and chunking), and the BioLemmatizer (Liu et al., 2012) are used for linguistic pre-processing, and are wrapped as UIMA Annotators. To enable the "Pattern Detection" step, we have developed a library (libfsmg) that provides generic Java classes for compiling non-deterministic finite state machines using backtracking for the identification of subgroup matches[5]. We then designed a sentence-level, (regular) syntax expression grammar that makes use of this library (see Table 2, bottom). The syntax patterns for this step are manually created from a "seed" collection of patterns (see Table 2, top), enriching them with linguistic properties using the displayed grammar (bottom). Terms in the generic patterns (top) shown in parenthesis are optional, but increase the specificity of a match, separating it farther from GREs and protein interactions events. In the case of **site** in the noun phrase head of a TF mention, the keyword may be replaced with "motif", "element", or "sequence". For **promoter**, also "promoter region", "enhancer", and "silencer" may be applied. The TF/TG body of any noun phrase may be used as prepositional complement ("TG be direct target of TF" instead of "TG be direct TF target"). For the special preposition within, the TG promoter element does not need to be present: For all other prepositions, we detected ambiguity with protein interactions, but all others may be substituted with proper alternatives. *Regulate/regulation* can be replaced with any of its synonyms or coordinate terms, possibly indicating the directionality of the event (e.g., "activate" or "inhibition"). In other words, directionality is determined by the presence of selected verbs (or their nominalized forms). For example, the generic pattern "TF direct regulate TG" can be re-written as "direct repression of TG by TF", indicating a down-regulation event.

To translate these patterns into syntax expressions, the LHS grammar rules (Table 2, bottom) in quotes or rectangle brackets represent terminals; `<token>` (describing a token) and `<lemma>` (describing the lemma of a token) can be substituted with any standard regular expressions. I.e., nested within the syntax expressions, regular expressions can be applied to match tokens and/or their lemmata. A Token can be quantified or generalized with a match-any expression (dot, "`.`"). Phrases (delimited by "`[`" and "`]`") can be declared optional ("`?`"). This is an example syntax expression that would implement "TF direct *regulate* TG" and detects an up-regulation event ("activate") of a TF on a TG (subgroups, delimited by "`(`" and "`)`"):

```
[ NP ( .  + ) ] .  * RB_direct [ VP . *
activate ] [ NP DT_* ?  ( .  + ) ]
```

This would match "Here, we show that intracellular Aβ42 directly binds and activates the p53 promoter, ..." (PMID 15548589) and produces the tokens "intracellular Aβ42" (TF) and "p53 promoter" (TG) as subgroup matches. The high specificity of these patterns also explains the need for mining article bodies, making it more likely to find instances of these restrictive patterns.

After the TRE pattern detection step, UIMA Annotators are provided for detecting lists of bio-entity terms (such as organism names or genes and protein symbols) in the two last steps, entity recognition and gene/protein mapping. For the first, the pipeline makes use of the Linnaeus Tagger (Gerner et al., 2010), a system specifically tailored for annotating term collections ("Gazetteers") of bio-entities, and in particular, provides a dictionary for matching organism mentions. However, the underlying finite state automaton[6] was not able to compile a state machine over the collection of approximately 23 million gene symbols collected for the `txtfnnl` pipeline. Therefore, instead, we implemented a concurrent PATRICIA trie[7]-based UIMA Annotator for handling very large collections (i.e., millions) of terms. Optionally, term matching in this trie-based Annotator can be limited to coin-

[3] `http://github.com/fnl/txtfnnl`

[4] `http://opennlp.apache.org/`

[5] `http://github.com/fnl/libfsmg`

[6] brics automaton, `http://www.brics.dk/automaton`

[7] `http://code.google.com/p/concurrent-trees`

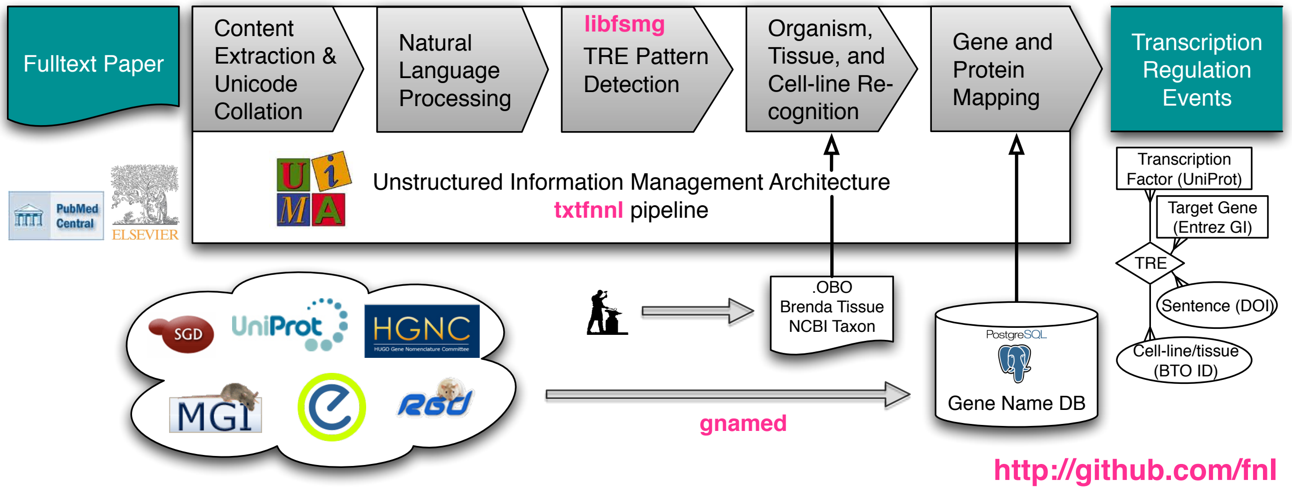
Fulltext Paper
Content Extraction & Unicode Collation
Natural Language Processing
libfsmg
TRE Pattern Detection
Organism, Tissue, and Cell-line Re-cognition
Gene and Protein Mapping
Transcription Regulation Events
PubMed Central
ELSEVIER
Unstructured Information Management Architecture
txtfnnl pipeline
Transcription Factor (UniProt)
Target Gene (Entrez GI)
TRE
Sentence (DOI)
Cell-line/tissue (BTO ID)
.OBO
Brenda Tissue
NCBI Taxon
PostgreSQL
Gene Name DB
SGD
UniProt
HGNC
MGI
RGD
gnamed
http://github.com/fnl

Figure 2: **The particular configuration of the UIMA-based `txtfnnl` text mining pipeline for transcription regulation event extraction.** The pipeline consist of a series of UIMA annotators; The first extracts content from most filetypes (e.g., PubMed Central XML, Elsevier XML, MEDLINE abstracts, or plain HTML), wrapping a modified Apache Tika (`http://tika.apache.org/`). Several Natural Language Processing (NLP) Annotators apply sentence segmentation, part-of-speech tagging, phrase chunking, and lemmatization. The pattern annotator (based on the `libfsmg` library) detects sentences that match a particular TRE pattern (see Table 2). Finally, several Annotators tag named entities with concept identifiers (from ontologies or databases) of particular organisms, cell-lines and tissue types, and gene or protein identifiers (that are first collected using the `gnamed` tool). This produces (potential) TREs consisting of the TF, TG, the sentence(s) the TRE is mentioned in, as well as any cell lines or tissues mentioned. The pipeline itself is open source, available via GitHub, and can be installed using Apache Maven. Because the patterns and Gazetteers are fed into it as parameterized arguments, the pipeline can be applied to other tasks as well.

cide with token boundaries, where this boundary is defined as the offset where the Unicode category of consecutive characters changes, for example, when a upper-case letter character is followed by a digit. The only exception to this boundary rule is a single upper-case letter followed by lower-case letters, as is the case for capitalized words. In addition, a variable number of separator characters (such as spaces, dashes, slashes, etc.) may be allowed between tokens, exact letter case matching can be required or disabled, and Greek letters may be collated to their Latin representations, all of which can be important aspects particularly when matching gene names and symbols. The gene names and symbols are provided via `gnamed`[8], a Python-based pipeline for bootstrapping a united PostgreSQL repository of symbols, names, and keywords per gene and across several sequence DBs. To make this gene normalization step more efficient, only the TF/TG subgroups matched by the TRE patterns are scanned by this Annotator.

5. Resources for TRE Mining and Curation

In addition to the pipeline itself, several other resources are under construction: (1) a compilation of mammalian transcription factors ("TFCheckpoint", (Chawla et al., 2013)), (2) an expansion to the PSI:MI ontology terms to cover experimental methods for TREs, (3) a collection of TRE-relevant patterns, (4) a Gold Standard of TRE-annotated articles, and (5) curation guidelines to create this Gold Standard.

The only pre-existing data that records functional TRE interactions on an per-article basis are the resources made available by transcription factor databases as, for example, TRED (Jiang et al., 2007) or ORegAnno (Griffith et al., 2008). However, these databases exclusively curate interactions for which relevant experimental evidence was generated in that article. In other words, if a direct, functional TRE is described only with cited evidence for the interaction, it will not be curated. Another important aspect to consider stems from the frequent use of supplementary or any other "external data" by human curators, such as following references to determine the correct sequence record to annotate. These issues imply that the use of commonly available bio-curation results as a Gold Standard for text mining systems will incur false positives for results that are not wrong per se, and will increase the false negatives for pairs that are impossible to extract ("external" material), leading to a possibly overly strict comparison and evaluation. While there do exist well known annotated corpora for GREs

[8] `http://github.com/fnl/gnamed`

TF bind to TG **promoter**
TF binding to TG **promoter**
TF **site** in TG **promoter**
TF bind within TG
TG be direct TF (transcriptional) target (gene)
TG (**promoter**) TF **site**
TG **promoter** bind TF
TF be recruit to TG **promoter**
TG (**promoter**) contain TF **site**
TF associate with TG **promoter**
TF interact with TG **promoter**
TF direct *regulate* TG
TF binding within TG
TG be direct *regulate* by TF
TF **site** within TG
TG *regulation* through TF **site**
TF recognize TG **promoter**
TF *regulate* TG experiment*

S	→	Phrase S? | Capture S? | Token S?
Phrase	→	“[“ Chunk InPhrase “]” “?”?
Capture	→	“(“ S “)”
InPhrase	→	CapInPhr InPhrase? | Token InPhrase?
CapInPhr	→	“(“ InPhrase “)”
Chunk	→	“NP” | “VP” | “PP” | “ADVP” | ...
Token	→	“.” Quantifier? | RegEx Quantifier?
Quantifier	→	“*” | “?” | “+”
RegEx	→	[[<token> “_”]? Tag “_”]? <lemma>
Tag	→	“JJ” | “NN” | “VBZ” | ...

Table 2: **The generic patterns judged by human experts to likely describe TREs (top). A syntax expression grammar (bottom) is then used to endow these generic patterns with linguistic context.** The generic patterns (top) are ordered by the frequency of the first matching passage extracted from TRED full text articles and are being manually translated to the syntax expressions (bottom). See text for details.

- such as the GeneReg corpus (Buyko et al., 2010), or the LLL05 corpus (Nédellec, 2005), in addition to the already mentioned GENIA Shared Task corpora - these collections were not created specifically for (direct) transcription regulation events. For all these reasons, we are now creating a dedicated Gold Standard where, on one hand, cited interactions may be included if the description allows biologists to deduce a direct TRE (implicitly trusting authors are reporting true facts), while at the same time “external data” is never used to create annotations. A good Gold Standard should represent an as diverse as possible sample; We selected 70 articles since 1999 onwards from 33 different journals and describing TRE interactions for over 100 different transcription factors from all major TF families (as defined by (Vaquerizas et al., 2009)). Furthermore, we have created curation guidelines to establish a common curation standard for extracting these interactions. We are currently in the process of annotating these articles each with at least three curators (trained in molecular biology), and will then keep resolving agreement issues until a final consensus is reached that overlaps with the guidelines. The final Gold Standard should contain, for each TRE, the TF as UniProt accession and the TG as EntrezGene ID. In addition, two lists may be annotated: the relevant tissue/cell type(s) (as BRENDA Tissue Ontology IDs) and the experimental evidence code(s) that were used to trace the interaction (as PSI Molecular Interaction ontology IDs). However, the IMEx’ PSI MI ontology focusses on detection of protein interactions rather than transcription (regulation) events (Orchard et al., 2012). Therefore, we are collaborating with members of the IMEx consortium to explore ways of expanding the PSI MI ontology with relevant experimental evidence types for detecting TREs that are not (yet) part of this ontology.

To generate the syntax expressions for the text mining pipeline, we extracted and lemmatized clauses that contained both a TF and TG mention from 1932 TRED-annotated articles. With the use of the MyMiner biocuration tool (Salgado et al., 2012), the 2756 clauses that appeared at least three times in the whole corpus were repeatedly classified by four independent expert curators as describing a direct TRE or not, until a consensus was reached for all of them, identifying 149 of the 2756 passages as describing direct TREs. The most generic patterns that cover these 149 cases are shown in Table 2 (top), while another 531 clauses, or roughly 3.5 times as many, were classified as describing GREs. All TRE clauses have some context-based reference in common that allowed the curator to infer a direct TRE; In the final patterns, this is ensured by either

1. The explicit mention of a *direct* transcription regulation event or TF *binding*.
2. The presence of the keyword *promoter* in the head of the noun phrase containing the TG.
3. The mention of a binding *site* in the head of the noun phrase containing the TF.
4. A reference to an experimental method that can be applied to detect a TRE.

These 18 patterns represent the exhaustive list of cases that permitted the curators to infer a direct TRE assuming the regulator is a *known* TF for all inspected 2756 clauses. The main conclusions drawn from this exercise are that a transcription-relevant DNA element or experimental method term must be present to allow inference of cis-regulatory events by experts. The only valid alternative is a traceable author statement declaring a *direct* promoter *binding* event (“TF directly regulates TG”). Finally, the regulator always must be a *known* TF. These strict limitations on the patterns represent the main difference of our approach to existing pattern-based systems. In another effort related to this project, (Chawla et al., 2013) have already created a repository of all known and putative mammalian (in particular, human, mouse, and rat) transcription factors and are now in the process of manually curating published experimental evidence that will allow classification of specific DNA-binding TFs. This work will be used to confine the extraction of the text mining pipeline by limiting the TF entity space that may be mapped for regulator mentions.

6. Prospects of this Work

The extraction system and resources presented explore and define standards for a public repository of transcription regulation events, similar to the existing MIMIx and IMEx standards for protein interactions (Orchard et al., 2007). The text mining aspects of the project provide a pipeline tweaked for high-precision to automatically extract TREs and bootstrap the low confidence repository, and a complementary high-recall version that will be used to assist human curation in the framework of a specific bio-curation tool developed for TRE extraction and annotation.

7. Acknowledgements

We would like to thank Karen Christie and Jane Lomax for introducing us to the transcription factor curation and ontology being developed at the JAX Laboratory and EBI. We are grateful to David Salgado for continuously updating MyMiner as to cater to our curation needs. FL wishes to acknowledge Miguel Vazquez particularly for the discussions on the algorithmic implementations of the pattern and entity matching annotators.

The work was supported by the grant BIO2007-66855 from the Spanish Ministry of Science and Innovation and by the Spanish National Bioinformatics Institute (INB-ISCIII).

References

S. Aerts, M. Haeussler, S. Van Vooren, et al. Text-mining assisted regulatory annotation. *Genome Biology*, 9(2):R31, 2008.

S. X. Atwood, M. Li, A. Lee, J. Y. Tang, and A. E. Oro. GLI activation by atypical protein kinase C [iota]/[lambda] regulates the growth of basal cell carcinomas. *Nature*, 494(7438):484–488, February 2013.

W. A. Baumgartner, K. B. Cohen, L. M. Fox, G. Acquaah-Mensah, and L. Hunter. Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics (Oxford, England)*, 23 (13):i41–8, July 2007.

E. Buyko, E. Beisswanger, and U. Hahn. The GeneReg corpus for gene expression regulation events: An overview of the corpus and its in-domain and out-of-domain interoperability. In *LREC 2010–Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 19–21, 2010.

A. Ceol, A. Chatr-Aryamontri, L. Licata, and G. Cesareni. Linking entries in protein interaction database to structured text: The FEBS Letters experiment. *FEBS Letters*, 582(8):1171–1177, April 2008.

K. Chawla, S. Tripathi, L. Thommesen, A. Lægreid, and M. Kuiper. Tfcheckpoint: a curated compendium of transcription factors. *Bioinformatics*, submitted, 2013.

S. Djebali, C. a. Davis, A. Merkel, et al. Landscape of transcription in human cells. *Nature News*, 489(7414):101–108, 2012.

K. Fundel, R. Küffner, and R. Zimmer. RelEx–relation extraction using dependency parse trees. *Bioinformatics (Oxford, England)*, 23(3):365–371, February 2007.

M. Gerner, G. Nenadic, and C. M. Bergman. Linnaeus: a species name identification system for biomedical literature. *BMC Bioinformatics*, 11(1):85, 2010.

M. B. Gerstein, A. Kundaje, M. Hariharan, et al. Architecture of the human regulatory network derived from ENCODE data. *Nature News*, 489(7414):91–100, September 2012.

O. Griffith, S. B. Montgomery, B. Bernier, et al. Oreganno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Research*, 36(suppl 1):D107–D113, 2008.

U. Hahn, K. Tomanek, E. Buyko, J. Kim, and D. Rebholz-Schuhmann. How feasible and robust is the automatic extraction of gene regulation events?: a cross-method evaluation under lab and real-life conditions. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pages 37–45. Association for Computational Linguistics, 2009. ISBN 1932432302.

S.-s. C. Huang, D. C. Clarke, S. J. C. Gosline, et al. Linking Proteomic and Transcriptional Data through the Interactome and Epigenome Reveals a Map of Oncogene-induced Signaling. *PLoS computational biology*, 9(2):e1002887, February 2013.

G. M. James, C. Sabatti, N. Zhou, and J. Zhu. Sparse regulatory networks. *The annals of applied statistics*, 4(2):663, 2010.

C. Jiang, Z. Xuan, F. Zhao, and M. Zhang. Tred: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Research*, 35(suppl 1):D137–D140, 2007.

J.-D. Kim, N. Nguyen, Y. Wang, et al. The Genia Event and Protein Coreference tasks of the BioNLP Shared Task 2011. *BMC Bioinformatics*, 13(Suppl 11):S1, June 2012.

R. Klinger, S. Riedel, and A. McCallum. Inter-Event Dependencies support Event Extraction from Biomedical Literature. In *Workshop on Mining Complex Entities from Network and Biomedical Data*, 2011.

M. Krallinger, C. Rodriguez-Penagos, A. Tendulkar, and A. Valencia. PLAN2L: a web tool for integrated text mining and literature-based bioentity relation extraction. *Nucleic Acids Research*, June 2009.

M. Krallinger, F. Leitner, C. Rodriguez-Penagos, and A. Valencia. Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biology*, 9 Suppl 2:S4, 2008.

M. Krallinger, F. Leitner, M. Vazquez, et al. How to link ontologies and protein-protein interactions to literature: text-mining approaches and the BioCreative experience. *Database*, 2012: bas017, 2012.

F. Leitner, A. Chatr-Aryamontri, S. A. Mardis, et al. The FEBS Letters/BioCreative II.5 experiment: making biological information accessible. *Nature Biotechnology*, 28(9):897–899, 2010a.

F. Leitner, S. A. Mardis, M. Krallinger, et al. An Overview of BioCreative II.5. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, 7(3):385–399, June 2010b.

B. Lenhard, A. Sandelin, and P. Carninci. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Reviews Genetics*, 13(4):233–245, 2012.

M. Levine and R. Tjian. Transcription regulation and animal diversity. *Nature*, 424(6945):147–151, July 2003.

C. R. Lickwar, F. Mueller, S. E. Hanlon, J. G. McNally, and J. D. Lieb. Genome-wide protein–DNA binding dynamics suggest a molecular clutch for transcription factor function. *Nature*, 484 (7393):251–255, April 2012.

H. Liu, T. Christiansen, W. A. Baumgartner, and K. Verspoor. BioLemmatizer: a lemmatization tool for morphological processing of biomedical text. *Journal of Biomedical Semantics*, 3 (1):3, 2012.

A. Manine, E. Alphonse, and P. Bessières. Learning ontological rules to extract multiple relations of genic interactions from text. *International Journal of Medical Informatics*, April 2009.

C. Nédellec. Learning language in logic-genic interaction extraction challenge. *Proceedings of the Learning Language in Logic Workshop (LLL05)*, 4, 2005.

S. Orchard, L. Salwinski, S. Kerrien, et al. The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nature Biotechnology*, 25(8):894–898, August 2007.

S. Orchard, S. Kerrien, S. Abbani, et al. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nature Methods*, 9(4):345–350, March 2012.

H. Pan, L. Zuo, V. Choudhary, et al. Dragon TF Association Miner: a system for exploring transcription factor associations through text-mining. *Nucleic Acids Research*, 32(Web Server): W230–W234, July 2004.

P. J. Park. ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, September 2009.

S. Riedel, D. McClosky, M. Surdeanu, A. McCallum, and C. D. Manning. Model combination for event extraction in BioNLP 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 51–55. Association for Computational Linguistics, 2011. ISBN 1937284093.

C. Rodriguez-Penagos, H. Salgado, I. Martínez-Flores, and J. Collado-Vides. Automatic reconstruction of a bacterial regulatory network using Natural Language Processing. *BMC Bioinformatics*, 8:293, 2007.

S. Roy, K. Heinrich, V. Phan, M. Berry, and R. Homayouni. Latent Semantic Indexing of PubMed abstracts for identification of transcription factor candidates from microarray derived gene sets. *BMC Bioinformatics*, 12(Suppl 10):S19, October 2011.

D. Salgado, M. Krallinger, M. Depaule, et al. Myminer: a web application for computer-assisted biocuration and text annotation. *Bioinformatics*, 28(17):2285–2287, 2012.

T. Sandmann, L. J. Jensen, J. S. Jakobsen, et al. A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development. *Developmental Cell*, 10(6):797–807, June 2006.

J. Saric, L. J. Jensen, R. Ouzounova, I. Rojas, and P. Bork. Extraction of regulatory gene/protein networks from Medline. *Bioinformatics (Oxford, England)*, 22(6):645–650, March 2006.

J. Saric, L. J. Jensen, R. Ouzounova, I. Rojas, and P. Bork. Extracting Regulatory Gene Expression Networks from PubMed. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

Y. Tsuruoka, Y. Tateishi, J.-D. Kim, et al. Developing a Robust Part-of-Speech Tagger for Biomedical Text. *Advances in Informatics - 10th Panhellenic Conference on Informatics*, 3746: 382–392, 2005.

A. Valouev, D. S. Johnson, A. Sundquist, et al. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods*, 5(9):829–834, August 2008.

J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics*, 10(4):252–263, April 2009.

H. C. Wang, Y. H. Chen, H. Y. Kao, and S. J. Tsai. Inference of transcriptional regulatory network by bootstrapping patterns. *Bioinformatics (Oxford, England)*, 27(10):1422–1428, May 2011.

H. Yang, G. Nenadic, and J. A. Keane. Identification of transcription factor contexts in literature using machine learning approaches. *BMC Bioinformatics*, 9(Suppl 3):S11, 2008.