BBA - Gene Regulatory Mechanisms 1865 (2022) 194778

Contents lists available at ScienceDirect

BBA - Gene Regulatory Mechanisms

journal homepage: www.elsevier.com/locate/bbagrm

ExTRI: Extraction of transcription regulation interactions from literature☆

Miguel Vazquez[a,*], Martin Krallinger[a], Florian Leitner[b], Martin Kuiper[c], Alfonso Valencia[a,d], Astrid Laegreid[e,*]

[a] Barcelona Supercomputing Center, Barcelona, Spain
[b] Data Catalytics, Madrid, Spain
[c] Department of Biology, Norwegian University of Science and Technology (NTNU), Trondheim, Norway
[d] ICREA, Barcelona, Spain
[e] Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology (NTNU), Trondheim 7491, Norway

ARTICLE INFO

Keywords:
Text-mining
Transcription factors
Gene regulation
Systems biology

ABSTRACT

The regulation of gene transcription by transcription factors is a fundamental biological process, yet the relations between transcription factors (TF) and their target genes (TG) are still only sparsely covered in databases. Text-mining tools can offer broad and complementary solutions to help locate and extract mentions of these biological relationships in articles. We have generated ExTRI, a knowledge graph of TF-TG relationships, by applying a high recall text-mining pipeline to MedLine abstracts identifying over 100,000 candidate sentences with TF-TG relations. Validation procedures indicated that about half of the candidate sentences contain true TF-TG relationships. Post-processing identified 53,000 high confidence sentences containing TF-TG relationships, with a cross-validation F1-score close to 75%. The resulting collection of TF-TG relationships covers 80% of the relations annotated in existing databases. It adds 11,000 other potential interactions, including relationships for ~100 TFs currently not in public TF-TG relation databases. The high confidence abstract sentences contribute 25,000 literature references not available from other resources and offer a wealth of direct pointers to functional aspects of the TF-TG interactions. Our compiled resource encompassing ExTRI together with publicly available resources delivers literature-derived TF-TG interactions for more than 900 of the 1500–1600 proteins considered to function as specific DNA binding TFs. The obtained result can be used by curators, for network analysis and modelling, for causal reasoning or knowledge graph mining approaches, or serve to benchmark text mining strategies.

1. Introduction

Regulation of gene transcription is one of the primary mechanisms by which a cell controls its response to internal and external stimuli involved in development, environmental changes or even pathogen infections. A gene is transcribed after a transcription initiation complex is formed. A crucial step in forming this complex involves DNA binding transcription factors (dbTF), which bind to specific regulatory sequences of the gene and recruit other factors around it [1]. A dbTF may also bind a regulatory sequence to inhibit transcription instead of promoting it. Establishing which dbTF regulates which specific genes and how allows the mapping of the cell regulatory network underlying cellular processes, as dbTFs constitute the link between signal transduction that initiates cellular decisions and the gene expression that carries them out [2].

The sequence specificity of dbTFs is a defining element in their relationship to target genes. Other transcription factors that do not bind DNA directly or bind DNA in a non-sequence-specific manner are also involved in transcription initiation complexes; however, these proteins do not play a major role in determining the target gene, so non-dbTF proteins were not addressed in this work. From now on, we shall reserve the term 'TF' to refer specifically to dbTFs and use the term 'co-TF' for transcription factors that do not bind directly to DNA. Furthermore, we use the term Transcription Regulation Interaction (TRI) to denote the regulatory relationship of a TF with its target gene (TG).

Systems biology exploits the fact that processes in biology can be

☆ This article is part of a Special Issue entitled: Curation of the Gene Regulatory Knowledge Commons edited by Dr. Colin Logie, Dr. Wyeth Wasserman and Dr. Julio Collado.

* Corresponding authors.
E-mail address: miguel.vazquez.g@bsc.es (M. Vazquez).

https://doi.org/10.1016/j.bbagrm.2021.194778
Received 7 May 2021; Received in revised form 22 November 2021; Accepted 29 November 2021
Available online 5 December 2021
1874-9399/

studied with the help of computational models that represent entities, such as genes and proteins, and their interactions or regulatory relationships, which allows their analysis in, e.g., signal transduction, complex formation, or transcription regulation. In practical terms, systems biology requires resources that describe this knowledge. These resources are constructed mainly through the work of biological curators (known as biocurators) and are based on information from experiments reported in the scientific literature and from experimental data files [3]. Literature curation is costly and slow and is carried out by a rather small number of highly specialized domain experts. Such curated knowledge is therefore extremely valuable but lags far behind and is not able to keep up with the rapidly growing literature [4]. Text-mining can be used to computationally extract relationships of interest, speed up curation by offering helpful curation stacks for curators, or even fill gaps in knowledge bases [5]. Therefore, the biocuration community has worked closely together with biological text miners in initiatives such as BioCreative to explore the use of natural language processing and machine learning approaches for improving or complementing more traditional manual literature curation [6].

The biomedical text mining community has been very active with respect to the extraction of relationships between bioentities from the literature, in particular for protein-protein interactions [7], gene-disease relations [8], chemicals-disease relations [9] or chemical-gene relations [10]. Despite its critical importance, transcriptional gene regulation relations have been less studied by the text mining community, even though the initial approaches on this topic date back more than 15 years [11]. These early attempts used relation extraction rules based on syntactic-semantic chunking as well as verbs referring to gene transcription (for activation, repression, regulation), resulting in the STRING text-mining system, which was adapted later by others for various gene regulation extraction tasks [12]. The use of text mining systems for the curation of regulatory annotations was promoted by the Open Regulatory Annotation Consortium [13]. They focused on the literature triage task, i.e., finding annotation-relevant publications and the extraction of sequence information with the aim of supporting the regulatory annotation process. Wang et al. [14] proposed bootstrapping strategies to generate description patterns to detect relationships between transcription factors and target genes, also covering patterns for active and passive voice sentences, but using an evaluation set limited to HIF-1 expert-tagged articles. Similarly, Tang and colleagues implemented a method called AutoPat, which relies on a set of seed patterns and unsupervised pattern learning to recover gene regulation relations from the literature [15]. The UIMA-based txtfnnl text mining pipeline for extraction of transcriptional regulation events was based on a manually crafted collection of seed patterns to generate syntax patterns enriched with linguistic properties [16]. The BioNLP shared-task community challenge ("GENIA task 2") covered general gene expression events defined as positive and negative regulatory relations [17]. For the TRRUST database, a text-mining system was used to produce sentences for manual curation exploiting MeSH descriptors, commonly used words for transcriptional regulation, dependency parse trees and rules [18,19]. The work we present here shows that despite all these efforts, only a fraction of TF-TG relations reported in literature have been considered by existing resources, and it suggests that renewed and extended text-mining assisted efforts are warranted to make existing TF-TG relation information more exhaustively available for biocuration and computational use.

Far fewer resources and text mining applications were developed for extracting TF-TG relations than for protein-protein interactions. Several aspects are particularly lacking, such as resources that aid manual curation or that relate high throughput experimental methods with relevant literature evidence and corpora that can help train and assess the quality of text mining systems.

ExTRI (Extraction of Transcription Regulation Interactions) aims to provide an exhaustive collection of putative TRIs extracted from PubMed abstracts. Our objectives are threefold: 1) to serve as a base for building productive TRI curation stacks, 2) to provide transcription regulation interaction (TRI) information with sufficient quality to be used in computational modelling or knowledge graph mining approaches to support causal reasoning, and 3) to serve as a reference for critical examination of the distribution of TRI information across literature and current knowledge bases. ExTRI provides an extensive catalogue of putative TRIs with a volume that goes much beyond what is currently available in public databases. The post-processing and quality control work performed on the results indicates the overall quality and use of the ExTRI resource.

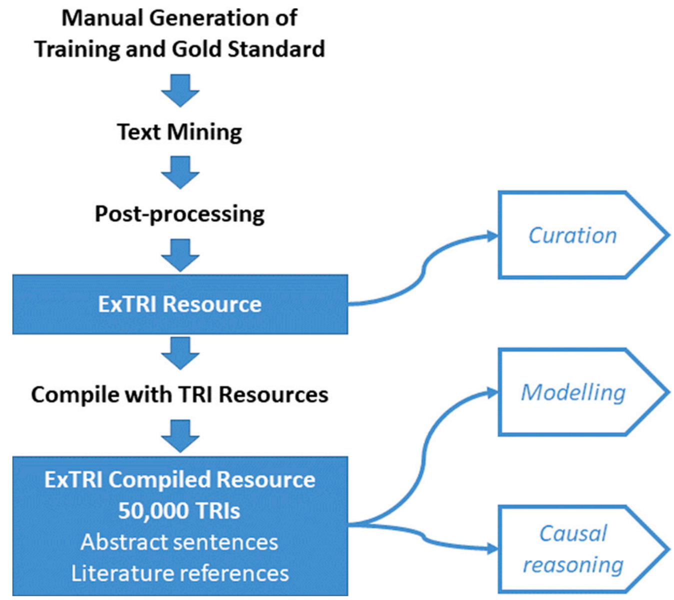

Fig. 1. Overview of the processes involved in generating ExTRI and the Compiled Resource combining TRI information from ExTRI with 'full confidence' manually curated TRI information from TFactS, GOA, IntAct, HTRIdb, SIGNOR, TRRUST, CytReg and GEREDB (see Table 2 for overview and references). The ExTRI resource is useful for building curation stacks. The compiled resource can be used as direct input for gene regulation network modelling and causal reasoning.

Text-mining has the obvious benefit of being able to quickly produce abundant results. But this benefit is offset by inaccuracies, confusing cases, accidental omissions, and other issues that might pass inadvertently. The quality of the ExTRI resource has been studied for a long time, involving several revision rounds, and it has been tested several times for its value in creating curation stacks throughout the years. In fact, the majority of the work presented here was done after the text-mining itself, which was performed some years ago and which therefore does not benefit from some of the most recent deep learning text-mining approaches, like Transformers (BERT) [20], and also does not cover articles published later than 2015. However, it is our belief that the efforts we have put into it, while they have consumed a long time, have helped to make the ExTRI resource much useful.

1.1. Overview of ExTRI

Fig. 1 outlines how the ExTRI resource was generated and how it was integrated and compared with other TRI data from existing public databases. The vast ExTRI resource of text mined abstract sentences, annotated with identified TF-TG relations, is useful for building curation stacks. Note that hereinafter we will often use the phrase 'text mined TRIs' or 'tm-TRIs' to remind the reader of the uncertainty inherent in the process that extracted these relationships. Users can also take advantage of the provided cross-references to the same TRIs in the other databases. The post-processing of ExTRI removes the most common errors and classifies tm-TRI sentences as high and low confidence so that they can

Table 1
Content of ExTRI corpus. Tm-TRIs - Text mined transcription regulation interactions. TFs - Specific DNA binding transcription factors. TGs - Target genes. TRI-sentences – Abstract sentences identified to contain TRIs. Abstracts – abstracts found to contain sentence(s) with TRI. Since some sentences support several TRIs, the number of unique sentences is lower.

ExTRI	All	High conf.
Tm-TRIs	40,453	18,437
TFs	991	865
TGs	5592	3848
TRI-sentences	94,185	52,862
Unique sentences	58,710	36,276
Abstracts	33,776	21,772

Table 2
Databases used for the compiled resource of TF-TG interactions. The table indicates whether all interactions or subsets of them were included. More details are given in Supplementary Document 5. Data were gathered from their online source during mid-2019.

Database	Content extracted for compilation	Reference
TFactS	All *(human, mouse, rat)*	Essaghir, 2010 [28]
HTRIdb	All *(human)*	Bovolenta, 2012 [6]
IntAct	Subset: protein-gene interactions *(human, mouse, rat)*	Kerrien, 2012 [29]
GOA	Subset: protein-gene regulatory interactions *(human, mouse, rat)*	Huntley, 2015 [30]
TRRUST	All *(human, mouse)*	Han, 2015; Han 2018 [19,31]
SIGNOR	Subset: interactions labelled with interaction mechanism 'transcriptional regulation' *(human, mouse, rat)*	Perfetto, 2016 [32]
CytReg	All *(human, mouse)*	Carrasco Pro, 2019 [33]
GEREDB	Subset: interactions with regulator TFClass TF *(human)*	Huang, 2019 [34]

be of a better use for gene regulation network modelling, either on its own or in combination with the compiled TF-TG resource.

2. Results

2.1. The ExTRI resource

The ExTRI resource was generated by text-mining assisted large scale extraction of information about regulatory interactions between specific DNA binding transcription factors and their target genes from MedLine abstracts. We focused our literature information extraction on the identification of gene and protein names, but this information, in general, is not sufficient to identify the species for which TRIs are reported. To address this, all the information is projected onto *Homo sapiens*, meaning that we consider either gene and protein mentions that can be directly matched to human gene symbols or that can be translated into human gene symbols using a cross-species lexical resource built by merging nomenclature resources from human, rat and mouse.

To improve the usefulness of the resource, we identified some common protein/gene name normalization errors that could be solved by post-processing rules. These rules were used to filter out erroneous TRI-classifications that were in most cases due to incorrect gene mention normalizations. We also built a validation dataset that we used to assess the initial quality of the ExTRI resources and to further classify all TRI sentences into high and low confidence using a machine learning classifier. This validation dataset is composed of 1500 sentences taken randomly from the ExTRI resource and was subject to two rounds of manual curation by two curators, both experts in the field of TRI validation. After an independent assessment, discrepancies between their labels were resolved by consensus. This dataset shows that about 54% of

Table 3
Comparison of ExTRI tm-TRIs with TRI from curated resources. The table shows counts of TF-TG pairs corresponding to these TRI, and the TFs and TGs for each of the databases, considering only data involving TFClass TFs. HC stands for High confidence, according to the databases criteria (some databases are considered entirely high confidence). The suffix -U refers to the counts for unique or exclusive entities (TF, TG, or TRI) for that database; in other words, present only in that database and not in the others. Counts for unique/exclusive entities for high confidence entries consider only high confidence entries in other databases as well. The data were filtered to retain only TRIs involving TFs that are members of TFClass (Wingender 2015, 2018). The filter was applied to make the data fully comparable since the proportion of data relating to non-TFClass TFs varies among the databases. This TFClass filter lowers the counts slightly, which explains the small discrepancies that can be seen between the results on this Table and Table 1,

Database	TRI All	TRI HC	TF All	TF HC	TG All	TG HC
ExTRI	38,620	17,824	903	785	5515	3757
ExTRI-U	29,051	10,964	155	98	155	1030
TFactS	6460	5651	308	308	2455	2038
TFactS-U	2727	2456	0	2	0	322
HTRI	2314	2314	265	265	1414	1414
HTRI-U	923	1022	9	14	9	462
Intact	242	242	82	82	128	128
Intact-U	141	150	0	1	0	27
GOA	898	898	410	410	580	580
GOA-U	369	420	10	15	10	67
TRRUST	7421	7421	529	529	2305	2305
TRRUST-U	2022	2532	2	8	2	248
Signor	797	797	224	224	443	443
Signor-U	148	179	0	0	0	13
CytReg	1283	1283	202	202	106	106
CytReg-U	387	490	0	0	0	15
GEREDB	7865	7865	630	630	2466	2466
GEREDB-U	3150	4073	11	30	11	326
Union	**49,903**	**31,163**	**944**	**887**	**6976**	**5731**

all ExTRI entries are correct (95% confidence interval: 51% to 57%). Cross-validation indicated that sentences classified as high confidence by our machine learning system were correct about 75% of the time.

The ExTRI resource contains approximately 40,000 different tm-TRIs represented in some 100,000 sentences obtained from about 34,000 abstracts (58,000 unique sentences, some sentences contain more than one unique TRI; Supplementary Table 1). Almost half of all TRIs are deemed high confidence by our classifier. These ~18,000 high confidence TRIs are derived from 41,000 unique sentences, which come from ~22,000 abstracts in the ExTRI set. The TRIs in our results feature almost a thousand TFs (15% less in high confidence TRIs only) and about 5600 TGs (20% less in high confidence TRIs only). See Table 1 for exact numbers.

2.2. Compiling ExTRI with other TRI resources

To assess the potential value of our ExTRI resource and its possible contribution to the current state of knowledge bases, we compared our resource of tm-TRIs with several databases containing TRI information (see Table 2 for databases and their references), focusing on their high confidence TRIs only: all TRIs derived from GOA, IntAct, TRRUST, CytReg, GEREDB and SIGNOR were considered to be of 'high confidence' (or rather 'full confidence', since they were manually curated); for TFactS and HTRIdb we set confidence cut-offs to select for annotations based on small- and mid-scale experiments. For TFactS, any TRI that originated from one unique PMID associated with more than 10 TRIs was considered low confidence, as their single evidence PMID likely represents a high throughput study. Likewise, for HTRIdb, all TRIs supported only by high throughput methods were considered low confidence.

To be able to compare the different resources with ExTRI, we normalized all the identifiers to standard HGCN using a translation table obtained from Ensembl BioMart, to which we added Entrez synonyms.

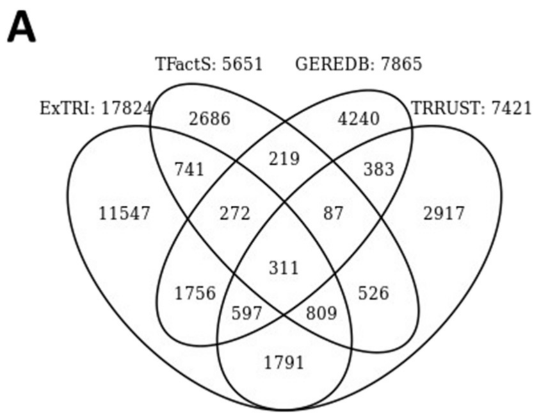

Fig. 2. Comparison of the main TRI resources. Overlaps of TF-TG interactions (panel A) and TFs (panel B) across ExTRI, TFactS, GEREDB and TRRUST, the four resources with the largest content, considering only high confidence TRI annotations for TFClass TFs, and excluding TF-TF autoregulation.

When looking for matches of ExTRI entries in databases, different gene name mentions that could both be a synonym for an official gene name were both considered as valid matches. The cases of AP1 and NFκB are special. Any mention of the names FOS* and JUN* (with * indicating specific family members) as TF was considered a valid AP1, and analogously, any mention of NFκB*, RELA or RELB as TF was valid for NFκB. For example, an ExTRI pair ‘AP1:<*Target gene*>’ would be considered a match if it was found as ‘FOSL1:<*Target gene*>’ in a database.

The ExTRI resource compiled with information on the other TRI resources is given in two tables: a *sentence coverage* overview (Supplementary Table 1) and a *TRI coverage* overview (Supplementary Table 2). The content of these tables is further elaborated in the section “Availability and usage”.

2.3. Contributions from each resource

For each database, we checked what information is unique for this database and not covered by any other database. Table 3 shows the counts for individual TRIs, TFs, and TGs in each database, after filtering for data involving only TFs that are members of TFClass [21,22]. This TFClass filter lowers the counts slightly but was applied to make the data fully comparable since the proportion of data relating to non-TFClass TFs varies among the databases. The union of almost 50,000 TRIs in this compiled resource contains roughly 31,000 TRIs that are either found in manually curated databases or with high confidence in the ExTRI resource. This comparison of all available information in this compiled resource shows the potential contribution of the ExTRI resource (see the second row from the top, labelled “ExTRI-U”). Thus, ExTRI could suggest roughly 11,000 (35%) tm-TRIs to the available manually curated TRIs, as these are not described yet in any of the databases in our compiled set. Also, ExTRI contains by far the largest number of transcription factors, including 98 TFs in high confidence TRIs involving a TFClass TF and not represented in manually curated TRIs in public databases.

It is of interest to note that for each database, roughly half of the manually curated TRIs are unique to that resource. Likewise, roughly half of the TRIs suggested in the ExTRI corpus are not found in the databases. This demonstrates that each of the resources can provide valuable contributions to the compiled resource of TRI information. The findings of the ExTRI indicate that a considerable volume of literature-documented knowledge concerning TRIs is still not recorded in manually curated public databases.

Exploring the overlaps between the four largest resources: ExTRI, TFactS, TRRUST and GEREDB, in terms of TRIs and TFs (Fig. 2), we see that only 311 TRIs and 268 TFs are listed by all of them. Of the 6277 TRIs that ExTRI has in common with at least one of the other three large resources, 3508 overlap with TRRUST and 2936 overlap with GEREDB.

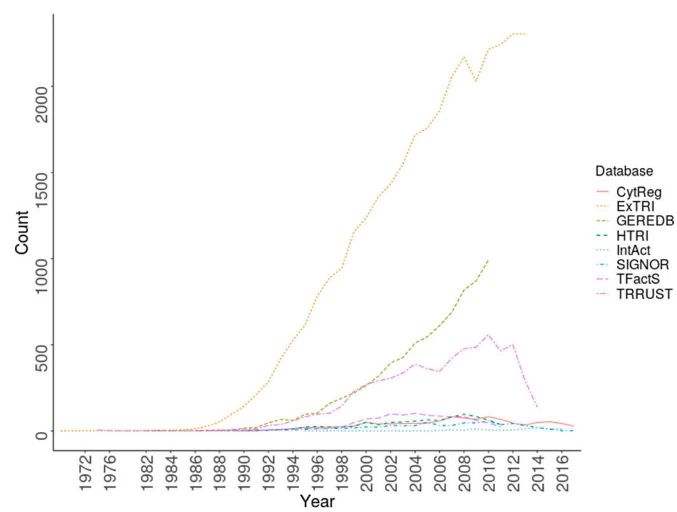

Fig. 3. Recorded TRI- papers per year. Total number of papers from which TF-TG interactions were recorded per year, across the different resources.

Additional Venn diagrams depicting the overlap between databases in the compiled resource are shown in Supplementary Document 1.

Examining their year of publication reveals that the number of papers from which TRIs have been text mined or manually curated rises sharply from the mid-1990-ties and appears to level out around 2010, at about 2300 papers per year for ExTRI and about 500 per year for TRRUST, while GEREDB TRIs rise to a maximum of 1000 papers for 2010 (Fig. 3).

2.4. Literature account for TF functionality covers only a fraction of known TFs and is unevenly distributed across TF families

The scientific focus on different transcription factors, as reflected in the literature, has been uneven, with some transcription factor families and specific family members being studied much more than others. We have analyzed this distribution, shown in Fig. 4, assuming that ExTRI, together with the other databases in our compiled resource (Table 3), represents a near exhaustive account of all TFs for which functional information is reported in the literature.

Our compiled resource lists tm-TRIs for a total of 944 TFClass transcription factors (Table 3, column “TF all”, last row labelled “Union”). When restricting to TFs that appear in at least one High Confidence TRI across all resources, this number goes down to 887. Thus, it seems that only half of the 1547 TFClass TFs have been reported in literature

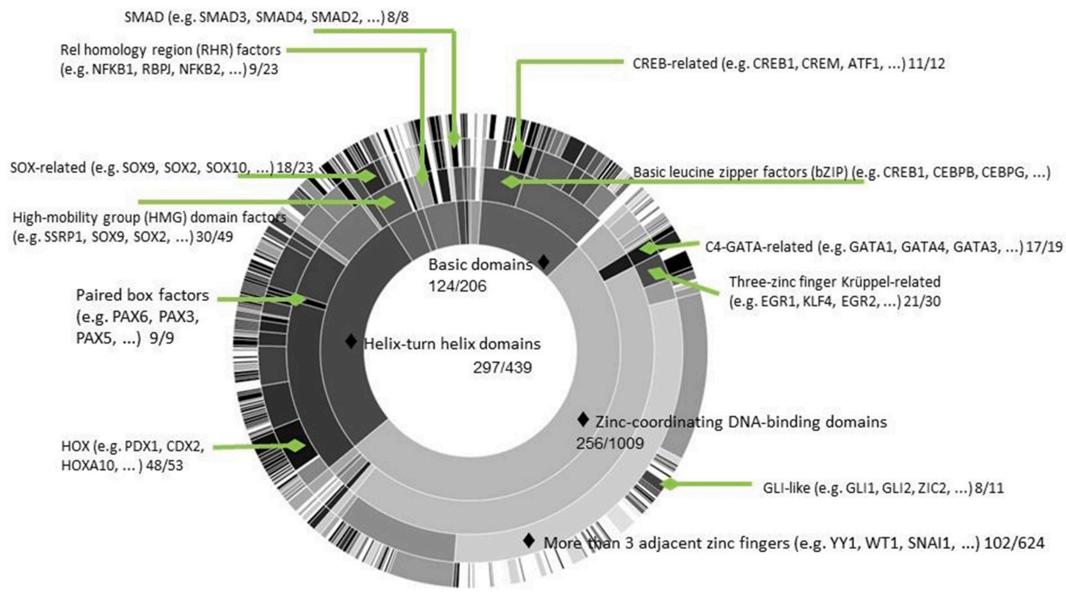

Fig. 4. Sunburst representation of TFClass TF. The image shows the distribution across TFClass families (Wingender et al., 2015, 2017) of TFs in TF-TG interactions (TRI and CTRI, all resources, Table 3). The TFClass family hierarchy starts with top classes in the inner circle (the three largest families (Basic domains, Helix-turn helix domains, Zinc-coordinating DNA binding domains) are marked). Subfamilies have increasing granularity from the centre to more outward circles. The area of each sunburst sector is proportional to the percentage of its TFClass family members in relation to all TFs. Darker shades represent TFClass families with a higher representation of TF family members in TF-TG database interactions. Some of the subfamilies are specified (green arrows), including how many of its members are cited in the databases (numbers in brackets) and with examples of TF family members given. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

abstracts with some information relating to their functionality in regulating the transcription of an identified TG.

As shown in Fig. 4, it is evident that some TFClass families have been quite thoroughly studied, in the sense that for very many of their protein members, literature holds information about target genes regulated by them. This applies for upper-level families like *Basic domains* and *Helix-turn-helix domains*, as well as for subfamilies like *High mobility group factors* and *Basic leucine zipper factors* and *Hox*. The low literature coverage of proteins in the large subfamily of *Zinc-coordinating DNA binding domains*, and in particular the family *More than 3 adjacent zinc fingers,* underpins the notion that a vast domain of potentially important gene regulatory information seems to have escaped experimental investigation so far. Two interactive versions of the sunburst displayed in Fig. 4 are available as Supplementary Images 1 and 2, one considering only the ExTRI resource and the other considering only the TRIs in manually curated public databases.

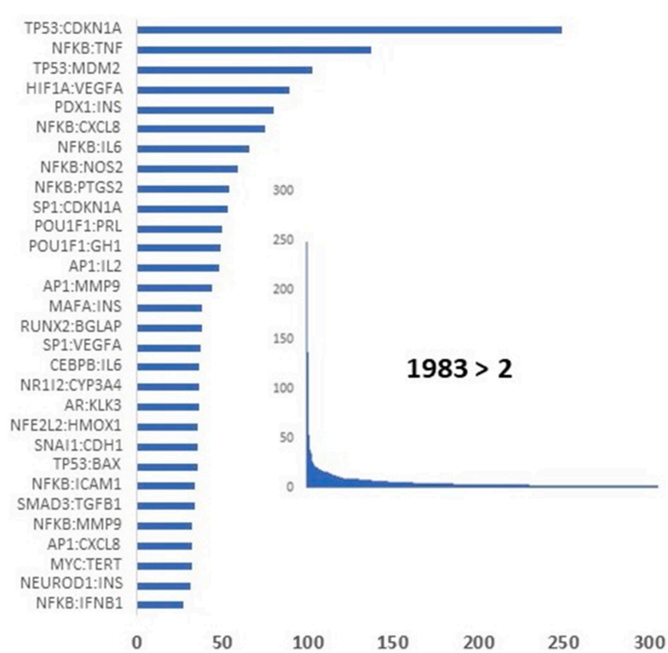

Fig. 5. Top 30 TF-TG pairs. Pairs are ranked according to the number of papers in which each of the pairs was found. The insert shows frequency distribution among the 1908 TF-TG pairs that were detected in at least three abstracts each. All data are from the TFClass, high confidence set, with autoregulation interactions not included.

ExTRI data indicate that close to 500 TFs are functionally characterized and reported in the literature in some detail, as judged by ExTRI listing at least five abstracts containing a high confidence tm-TRI for that TF. The majority of the ExTRI high confidence tm-TRIs (14,346) is extracted from only one or two abstracts. However, 4635 of these are also reported in at least one of the other TRI resources, indicating that they are deemed valid also by other managers of such knowledge. Thus, while data from fewer independent research papers may be available for these TRIs, they may nevertheless represent valid TF-TG functional information.

2.5. ExTRI greatly extends computational access to literature accounts of TF functionality

Considering TRIs in high confidence ExTRI sentences that match a manually curated TRIs in some database that have references to PMIDs (6477 TRIs in total), the majority of these (4850) have some overlap on their PMIDs, meaning that both ExTRI and manual curators have extracted identical TRIs from the same scientific papers. However, even for most of these TRIs, ExTRI contributes additional PMID references. For 4101 of the TRIs that are listed in several resources with the same papers as reference, ExTRI cites additional literature references, potentially extending convenient access to information about the biological context and evidence for these TRIs. ExTRI contributes in total 13,364 of these additional PMID references (with 5478 PMIDs mentioning information about more than one unique TRI). The 12,051 high confidence

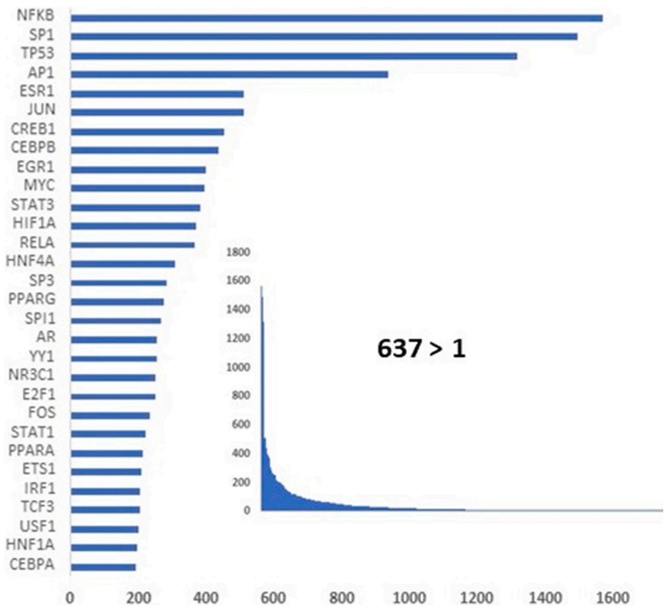

Fig. 6. Top 30 TFs. TFs are ranked according to the number of papers in which they are found as part of a TF-TG pair. The insert shows the frequency distribution among the 775 TFClass TFs represented in the high confidence set (autoregulation interactions not included).

tm-TRIs reported only by ExTRI and not found in the other resources are extracted from a total of 16,192 paper abstracts. Altogether, the high confidence ExTRI TRIs contribute 24,836 literature references not available from other resources.

It should be noted that the large set of 52,862 high confidence tm-TRI abstract sentences provided by ExTRI offer a wealth of direct pointers to a potentially wide set of functional aspects of the 18,437 TRIs stated therein, including sign (activation/repression), biological context (e.g., cell- or tissue type and state) and experimental evidence. This information can be directly used in curation processes and may, for many knowledge management purposes, contain sufficient evidence in the sentence itself, thus saving the user time-consuming work involved in reading a larger part of the text. As described in the section "Availability and usage", accessing the ExTRI-resource via BioGateway also gives users easy access to the full abstract from which the sentence is extracted.

2.6. *Prevalent TRIs, TFs and TGs*

We can assume that the frequencies with which transcription factors and target genes and their interactions are found in ExTRI reflect their representation in scientific literature, and thereby the attention they receive in scientific research. The TRIs most frequently observed in ExTRI include TP53-CDKN1A, NFκB-TNF, TP53-MDM2, HIF1A-VEGFA, PDX1-INS, NFκB-CXCL8 and NFκB-IL6 (Fig. 5). Each of these tm-TRIs was extracted from more than 60 abstracts.

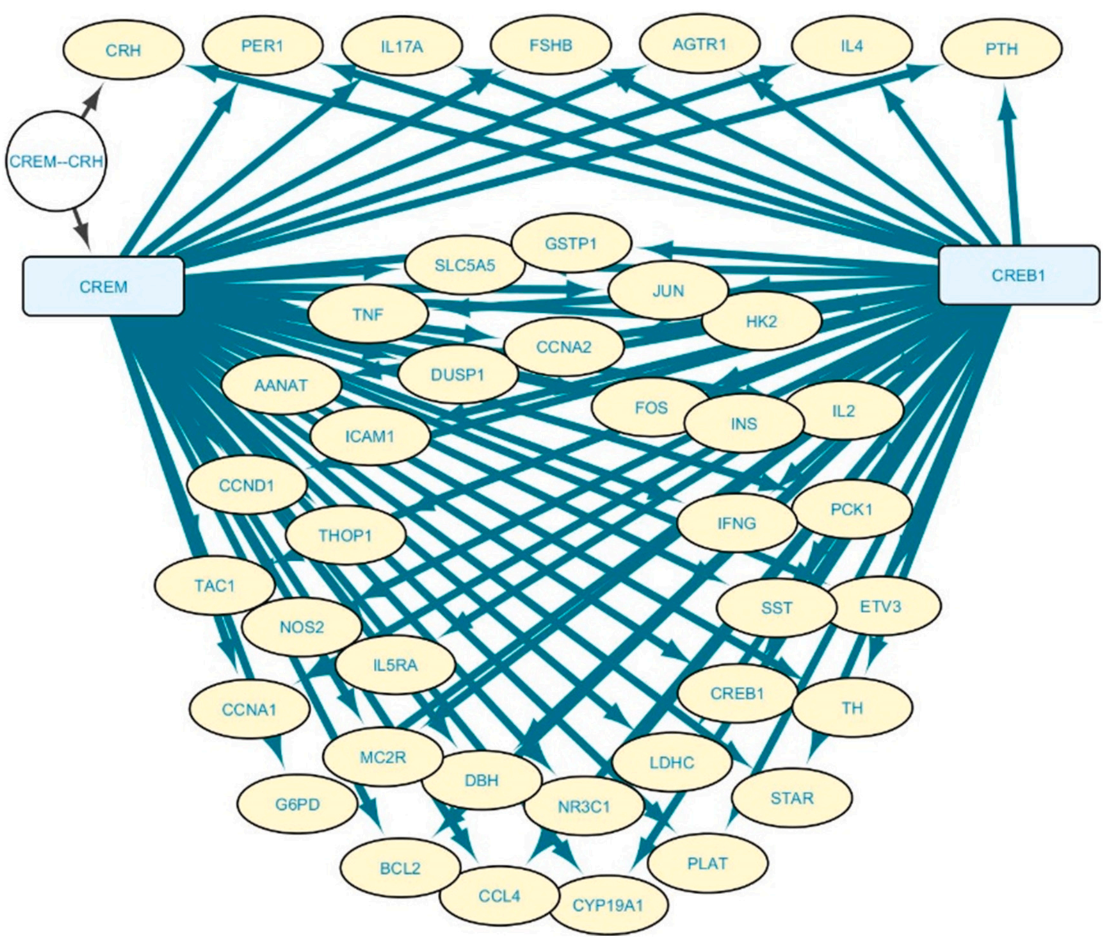

Fig. 7. CREB1-CREM coregulation network. A TF-TG network is shown with TGs (yellow ovals) targeted (green arrows) by both CREB1 and CREM (blue boxes) as extracted and displayed by the BioGateway app based on the compiled set of TF-TGs from all resources. Only TRIs supported by at least one High Confidence annotation are included. The white circle (top-left in the network) shows a node that results from expanding (double-clicking) a regulation arrow, which provides a link to the landing page for this TRI. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

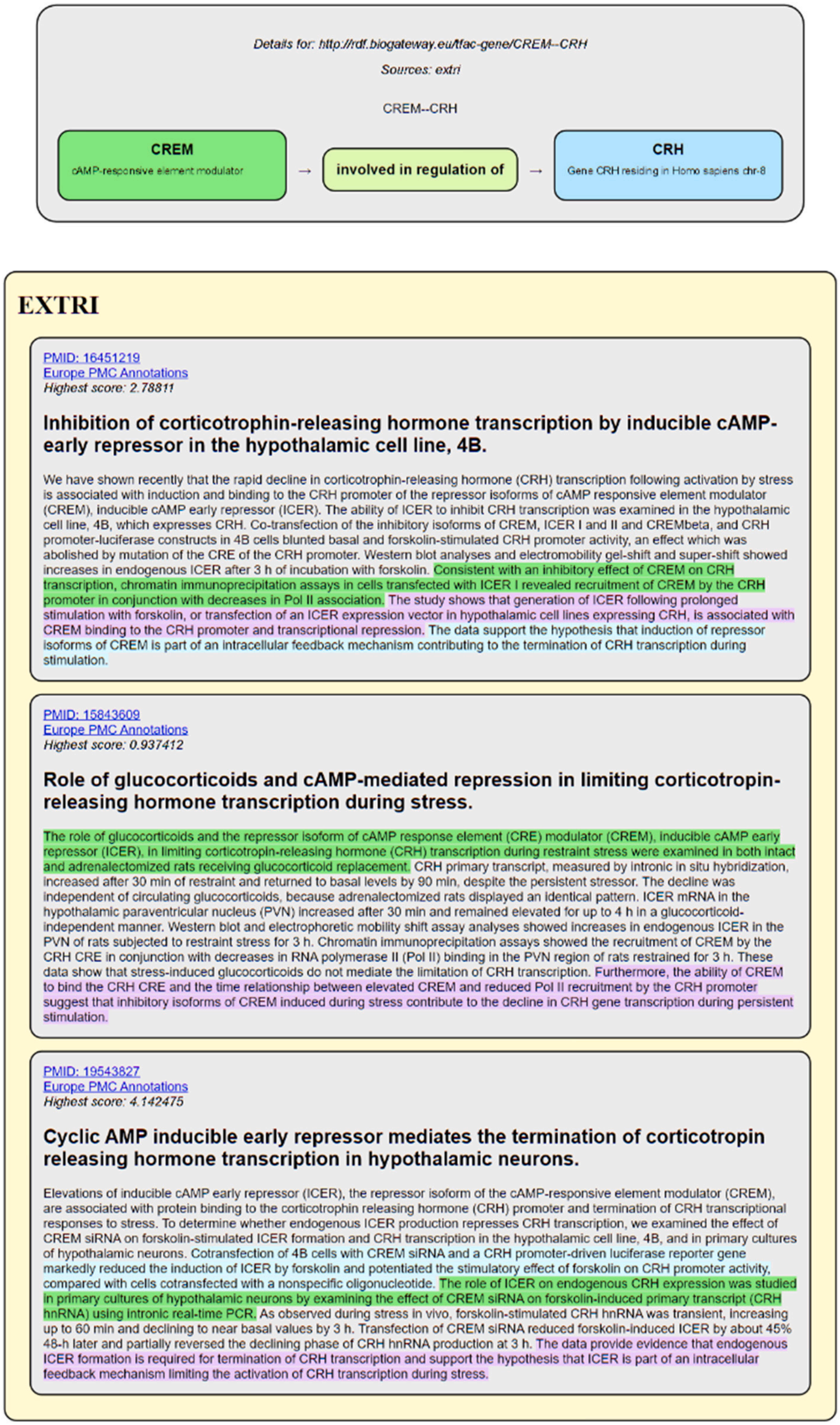

Fig. 8. TF-TG landing page. BioGateway view of abstracts with ExTRI-identified sentences containing information on CREM(ICER) regulation of CRH - Corticotropin-Releasing Hormone.

The two high-volume TRI databases in our compiled resource, TRRUST and GEREDB, both sample individual TRIs from multiple literature abstracts in curation strategies involving manual curation of abstracts identified by exhaustive text-mining of MEDLINE. The frequencies of individual pairs in TRRUST and ExTRI are similar, with an overlap of 12 TRIs among the top 30 in both resources (or 20 in the top 50). The top 30 TRIs in TRRUST are reported in 18 to 55 abstracts each, while in ExTRI, the top 30 tm-TRIs are detected in 28 to 262 abstracts each. The TRRUST resource fails to report on the high number of abstracts reporting on TRIs, including PDX1-INS, MAFA-INS, RUNX2-BGLAP, CEBP-IL6, and NEUROD1-INS, which all rank among the top 30 in ExTRI (each of these TRIs found in 30–75 abstracts), while they are

Table 4
AP1 and NF-κB dimer composition.

Identifier	AP1	NF-κB
Aliases	Activator protein 1 Transcription factor AP1	NF-κB Nuclear factor κ B Nuclear factor kappa-light-chain-enhancer of activated B cells
Dimers (HGNC symbols)	JUN-FOS, JUN-JUN, JUN-JUNB, JUN-JUND, JUN-FOSB, JUN-FOSL1, JUN-FOSL2 JUNB-JUNB, JUNB-JUND, JUNB-FOS, JUNB- FOSB, JUNB-FOSL1, JUNB-FOSL2, JUND-JUND, JUND-FOS, JUND-FOSB, JUND-FOSL1, JUND-FOSL2	NFKB1-RELA, NFKB1-NFKB1, NFKB2-NFKB2, NFKB2-RELB

referenced to only 2–5 abstracts each, in TRRUST. In GEREDB, the frequency ranks per TRI strongly differ from those found in ExTRI and TRRUST (only two of the ExTRI top 50 TRIs are also among the top 50 in GEREDB). This could be due to a manual curation strategy different from TRRUST, in the sense that TRIs might not necessarily have been selected for curation with a frequency proportional to their abundance in the text mined curation stack. Alternatively, the text mining strategy applied by GEREDB may have resulted in very different results in terms of frequencies compared to ExTRI and TRRUST. Further investigations would be necessary to establish whether ExTRI reflects the frequency of literature reports on individual TRIs in a more unbiased way than TRRUST and GEREDB.

Fig. 6 provides an overview of the most frequently extracted transcription factors, with NFκB, SP1 and TP53 retrieved from the highest numbers of papers (each in ~800–1500 papers). The remaining 26 TFs from the top 30 were found in 200–500 papers each (ESR1, JUN, CREB1, CEBPB, EGR1, MYC, STAT3, HIF1A, RELA, HNF4A, SP3, PPARG, SPI1, AR, YY1, NR3C1, E2F1, FOS, STAT1, PPARA, ETS1, IRF1, TCF3, USF1, HNF1A, and CEBPA). The ranking of highly studied transcription factors is similar in the union of all resources other than ExTRI (Supplementary Document 1), with 39 of the ExTRI top 50 also in the other top 50 (25 from top 30; 17 from top 20).

2.7. Availability and usage

The **Sentence coverage overview** (Supplementary Table 1) takes all ExTRI sentences and cross-references the TRI with entries identified with other databases, allowing the user to see additional evidence for each pair found in each sentence. The **TRI coverage overview** (Supplementary Table 2) presents TRI coverage across all databases, including the additional information provided by each database, The **Sentence coverage** file (Supplementary Table 1) can be used as a resource to guide curation, to examine the provenance of each reported TRI, or to gain access to some functional details about a particular TRI. The resource identifies the TRI level of confidence (high/low) as assessed by our machine learning approach. We have also incorporated all information about each TRI available from other resources as additional columns. The **TRI coverage overview** file (Supplementary Table 2) merges the tm-TRIs from ExTRI with the manually curated TRIs from public databases. ExTRI is treated as equivalent to the other resources in this table, which means that all instances of TF-TGs are listed, regardless of whether they come from a text mined TRI or a manually curated TRI. This resource can be used for modelling or for representing the information in a network or knowledge graph form. When presenting information from the curated resources, we have tried to retain as much information as possible (e.g., sign, confidence, or PMID for the article that sourced the annotation). Note that all the resources have been 'projected' into HGNC gene symbols to allow merging them. Mentions to AP1 and NFκB complexes were allowed to match also with their constituent proteins in accordance with dimer composition, as detailed in Table 4.

We have made the compiled TF-TG interaction resource available through the BioGateway database [23] (www.biogateway.eu). This database is a so-called triple store containing RDF graphs, and the ExTRI TRI resource, together with the TRI data from the other resources mentioned in this paper (Supplementary Table 2), was added as a TRI RDF graph. Although the data can be queried through the BioGateway endpoint[1] with the SPARQL query language,[2] the TRI data can more easily be assessed for network building in the Cytoscape environment through the BioGateway App [24] (see example in Fig. 7). In addition to the TRI data, the BioGateway resource provides a variety of Elixir Core Resource data for explorative network building, and the TRI instances can be added to ground annotated protein interaction subnetworks on transcriptional regulation. TRIs suggested for network building can be checked through link-outs to landing pages that display the origin of any TRI, and if they originate specifically from ExTRI, the sentence identified in the text mining exercise is highlighted in the abstract (see Fig. 8).

The example shown in Fig. 7 depicts the result obtained with a BioGateway App query (see Supplementary Document 2), retrieving a CREB1-CREM joint TF-TG regulatory network from BioGateway. The network of TGs regulated by these two TFs is of interest because it is well established that CREB1 is a predominantly activating transcription factor whose transcription regulatory activities (affecting a large set of target genes) is subject to negative feedback by the CREM gene splice variants ICER [25,26]. The expression of ICER proteins is itself under transcriptional activation control by CREB1. ICER proteins contain DNA binding domains but no transcription activating domains and can bind to similar transcription factor binding sites in the regulatory regions of target genes as CREB1. Thus, by competing with CREB1-binding, ICER represses many CREB1 target genes.

The BioGateway query (Supplementary Document 2) asking for genes that are connected to both the CREM and CREB1 transcription factor produces a TF-TG network with 42 target genes (BioGateway build 3.2.5), all supported by high confidence TRIs (see Fig. 7). Seven of these TGs (upper region of Fig. 7) are supported by CREB1 or CREM TRI information only available in ExTRI and not in any of the other resources: CREB1-AGTR1, CREM-AGTR1, CREM-CRH, CREB1-FSHB, CREM-IL17A, CREB1-IL4, CREM-PER1, CREM-PTH. All these TF-TGs, except for one (CREM-AGTR1), are identified from multiple sentences in two or three abstracts each. Validation of the abstract sentences identified by ExTRI shows that all except two (CREB1-IL4, CREM-PTH) are valid TRI statements. All sentences with their validation are shown in Supplementary Table 3. The interaction between CREM and CRH (Corticotropin-Releasing Hormone) (upper left corner, Fig. 7) is shown in an expanded form, by way of an interaction node through which BioGateway provides a link-out function to a landing page with information about the literature source of this tm-TRI. The CREM-CRH landing page is shown in Fig. 8, depicting the three MEDLINE abstracts with highlighted TRI sentences that are identified by ExTRI to hold evidence for the CREM-CRH relation. Each abstract also has a link to the original PubMed ID.

3. Discussion

The ExTRI resource vastly increases computationally accessible structured information concerning the involvement of specific DNA binding transcription factors in gene regulation. Our framework not only uses semantic and strict *functional class*-guided entity recognition to infer TRIs; it also recovers text fragments that can help assess their functional importance in specific cellular processes.

A comparison with existing TRI resources reveals that our automatically obtained interactions contribute significantly to the public structured knowledge about TRI. The interactions found by this work

[1] https://www.biogateway.eu/sparql-endpoint/.
[2] www.w3.org/TR/rdf-sparql-query.

can be used for generating models and knowledge graphs, from which TRIs can be easily validated in their abstract of origin and be used to generate experimental leads. Relationships deemed by our assessments to be of high confidence allow for a reasonable level of true positives. Moreover, the resource can be used for further information extraction efforts, not only by guiding manual curation but also in new text mining efforts. Overall, ExTRI significantly increases the coverage of computationally accessible information on transcription regulation interactions described in the literature.

Admittedly, projecting all the TRIs to human genes leads to false entries. However, most information in the literature regarding TRIs originates from studies on human, mouse and rat, for which ortholog genes often share a common spelling and for which many TRIs are conserved. Additionally, assuming that TRIs translate across this subset of mammals is common practice. These assumptions not only made the text-mining simpler, but we also feel it simplifies the use of the resource.

These limitations and assumptions, which are common to many text-mining applications, require some caution from users of the ExTRI resource. On the other hand, ExTRI's level of coverage and its reasonable level of quality in many settings provides definite advantages over the alternative of working with high-confidence but low coverage TRI annotations otherwise available in manually curated databases. As we will see later, ExTRI contributes target gene information annotations for some 100 TFs that were until now not listed with target genes in curated resources. Moreover, ExTRI also greatly extends the coverage and breadth of regulatory annotations for the ~900 TFs for which such information can be found in the literature. Finally, even for curated TRIs, ExTRI provides an extensive additional source of functional information, in the form of literature references with identified sentences from abstracts holding information not only about the specific TRI but also about the biological context, such as activation or repression, cell- or tissue type and state, and the experimental evidence underpinning the relation.

The availability through BioGateway, and more readily through the Cytoscape BioGateway App, provides the means to incorporate the compiled resource, including ExTRI and public databases, into the mainstream of biological network exploration. Although the number of false positives is rather substantial when using the unfiltered set of interactions, inevitable with current text-mining technologies, Cytoscape offers the possibility to alleviate this concern by enabling users to check the validity of any TRI by linking out to landing pages holding the abstracts and highlighted sentences that support them. In addition, validated sentences may serve in a feedback loop to increase and improve the text mining gold-standard in further rounds of text-mining.

Our analysis of the compiled TRI resource (Table 3) shows that approximately half of the content of each individual resource is unique for that specific resource. This indicates that a significant proportion of TF-TG regulatory interactions reported in the literature is not yet recorded in existing TRI knowledge bases. However, this analysis also suggests that our current compiled resource of TRIs, including ExTRI, may indicate a very significant gap in the space of possible TF-TG interactions: target genes that are regulated by the hundreds of TFs that seem not yet to be experimentally investigated as they are not found in literature report. This is most notable for TF members of the family of *Zinc-coordinating DNA binding domains*, and in particular for the subfamily *More than 3 adjacent zinc fingers* (Fig. 4), and it exemplifies the potentially vast range of transcription regulation interactions that still remain to be investigated.

The work presented here suggests that automated extraction of transcription regulation interaction information from text is a valuable approach to facilitate knowledge base curation, with a high confidence subset of results meeting a level of accuracy that might make it suitable for direct use in certain systems biology approaches. Importantly, ExTRI should be followed up with improved technology, including deep learning approaches. Such future work should benefit from The Minimum Information about a Molecular Interaction Causal Statement (MI2CAST) [27], suggesting a richer set of information that should be the aim of causal relationship annotation, such as sign (activation/repression), biological context (e.g. cell- or tissue type and state), or experimental evidence. This is especially true if the mining is extended to full texts, including tables and figures.

Automated information extraction procedures, as outlined above, will be applicable to extract information also for other transcription regulatory mechanisms than direct TF-TG interactions. This includes regulatory microRNA interactions, which are mentioned in the scientific literature in semantic terms and biological contexts very similar to interactions involving specific DNA binding TFs. Furthermore, molecular mechanisms involved in the interplay between dbTFs and co-TFs might also be captured by automated information extraction. Even if the results of text-mining are not 100% accurate, they can provide a wealth of information and pointers to functional aspects of TF-TG relationships that can be assessed by curators. Given the fact that current resources all seem to have tapped into partially complementary segments of the publication space, there is still much to be expected from new advanced text mining efforts.

4. Methods

4.1. Generation of training corpus and gold standard

The corpus used for the text-mining process was generated by carefully following a detailed curation process. The curation process was designed to produce a collection of valid TRI sentences from PubMed abstracts describing mammalian TFs having a direct regulatory relationship with one or more TGs, and in which the participants can be mapped unambiguously to unique database identifiers. Curation aimed to identify direct interactions between a mammalian DNA binding transcription factor, as provided by TFClass, and a target gene mediated by cis-regulatory DNA sequences of the TG. Curation specifically aimed to distinguish these TRIs from statements pertaining to indirect interactions, where the transcription regulatory effect is triggered by some upstream agent (for example, a receptor-ligand) or is exerted through some intermediate agent (for example, a signal transduction protein or co-TF). We regard TRIs to be a specific subclass of *gene regulatory interactions* (GRI), which comprise both direct TF-TG interactions as well as other interactions affecting the transcriptional regulation of a gene.

Each sentence was labelled according to curation guidelines that were drafted to describe the rules to standardize the identification of these binary relationship pairs in single sentences from scientific publications, and which enabled the curators to uniformly label direct interactions between a mammalian TF and a TG. Whenever it was possible to ascertain that the protein/gene mentioned referred to a species other than human, mouse or rat, the sentence was judged not eligible for the gold standard (GS) corpus. Descriptions of indirect GRIs were not labelled in the curation process. Due to this strict definition, the guidelines aim to distinguish between a potentially indirect GRI and its subclass TRI that specifically focuses on direct TF-TG interaction. The sentences collected by this process describe one or more such TRI pairs, as agreed by domain experts.

The manual curation followed four sequential steps: 1) Identification of sentences describing gene regulation interactions (GRI); such a sentence describes regulatory interaction between regulator and target; the regulatory interaction is not indirect, and semantic connection between regulator and target places the interaction within the context of transcription regulation; 2) Identification of regulator and target (Regulator is protein, Target identifies a gene and not a gene family or collection of genes); sentences that fulfil selection criteria in 1) and 2) are gene regulation sentences (GRI) with among them the subclass of transcription regulation sentences (TRI); 3) Normalization of the regulator to TF (TFClass); 4) Normalization of the target to gene. Sentences complying with steps 1 through 4 are positive TRI sentences. The final version of the curation guidelines covers a range of critical issues encountered

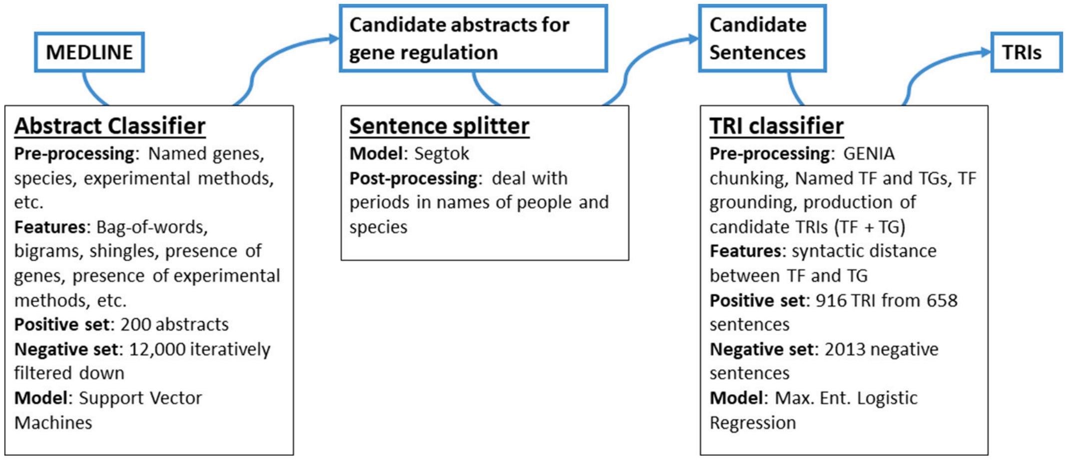

Fig. 9. Overview of text-mining approach to generate tm-TRIs. A collection of candidate abstracts was filtered from the entire set of abstracts in the MEDLINE 2015 baseline using a classifier. The abstracts were divided into sentences. Sentences with mentions to a TF and to another gene were further processed. Each pair of TF and gene were considered a candidate TRI (there could be several per sentence). A second classifier was used to investigate each candidate TRI present in each candidate sentence.

during this manual curation process, and it is available as Supplementary Document 3.

Each sentence in the GS was labelled independently by at least two domain experts. The annotations were compared, and in case of disagreements, decisions were discussed against the curation guidelines until consensus was reached. The GS is available as Supplementary Table 4.

The curation process was designed to capture only binary regulator-target pairs, while regulatory complex dependencies (e.g., a complex that requires the presence of several [co-]factors) were not annotated as such. Furthermore, TRI descriptions that were not expressed in single sentences, i.e., pairs that can only be established from the context of more than a single sentence of a publication, were not included in the corpus.

The final GS comprises 658 unique positive sentences from 189 different abstracts and specifies 916 TF relations between 208 different TFs and 208 different TGs (Full corpus is given in Supplementary Table 4). These positive and negative TRI sentences, selected and annotated with a list of binary interaction partners, constitute the central resource upon which the literature information extraction in the current work was built. Importantly, this GS is available to others to train their natural language processing systems and machine learning models.

4.1.1. Examples of positive TRI-sentences●●●●●●●●●●

Below we show some examples of positive TRI sentences from the GS (Supplementary Table 2). The TF-TG relationship is indicated with official HGNC symbols, and the PMID abstract source for the sentence is given in parentheses after each sentence.

- Bacterially expressed Tst-1 binds specifically to the promoter of the gene encoding myelin protein P0, a Schwann cell surface adhesion molecule. (POU3F1-MPZ; PMID: 1705013)
- The myeloid zinc finger gene, MZF-1, regulates the CD34 promoter in vitro. (MZF1-CD34; PMID:7579328)
- Regulatory role of MEF2D in serum induction of the c-jun promoter. (MEF2D-JUN; PMID:7760790)
- In the adult pituitary gland, Ptx1 appears to be recruited for cell-specific transcription of the POMC gene. (PITX1-POMC; PMID:8675014)
- Mutagenesis of each of the two sequence elements required for ILF binding decreased IL-2 promoter activity when assayed in transfection assays. (FOXK2-IL2; PMID:9065434)
- Furthermore, MEF2B overexpression in smooth muscle cells up-regulated the SMHC promoter, suggesting that MEF2B is important for SMHC gene regulation. (MEF2B-MYH11; PMID:9430690)
- Identification of Mad as a repressor of the human telomerase (hTERT) gene. (MXD1-TERT; PMID:10723141)
- The IL-8 promoter contains three ETS binding sites, and we identified the specific site that binds MEF and is required for MEF responsiveness. (ELF4-IL8; PMID: 14625302)
- The Krüppel-like zinc finger protein Glis3 directly and indirectly activates insulin gene transcription. (GLIS3-INS; PMID:19264802)
- FOXO1 stimulates ceruloplasmin promoter activity in human hepatoma cells treated with IL-6. (FOXO1-CP; PMID:21185807)

4.1.2. Examples of negative TRI-sentences

- In prostate cancer CWR22-Rv1 and PC3 cell lines, which showed distinct δ-catenin overexpression, E2F1 and Hes1 expression pattern was altered. (PMID:21106062)
- A significant positive correlation between the KLF5 and ZEB1 transcription levels in the pancreatic tumor tissues was observed. (PMID:30168064)
- In neuroendocrine prostate cancer mouse model derived allograft NE-10 tumors, δ-catenin showed an increased expression while Hes1 expression was diminished. (PMID:21106062).

4.3. Text-mining

An overview of the text-mining pipeline is shown in Fig. 9. The entire set of MEDLINE abstracts (about 24,000,000) up until 2015 was used (the 2015 MEDLINE baseline). The first step was to determine which of these abstracts were worth considering for our purposes. We trained a classifier that identified some 100,000 candidate abstracts, consisting of about 900,000 sentences in total. Our aim was to detect the TF-TG pairs for which the sentences describe a valid *transcription regulation interaction* (TRI). The details of the abstract classifier are provided in Supplementary Document 4. The gold standard used for abstract classification is available in Supplementary Table 5.

Sentences from the candidate abstracts were annotated with mentions of genes and proteins since these are the potential TFs and TGs in the TRI. In addition, the sentences were annotated with other entities (e.g., species and experimental methods) and *natural language processing* (NLP) features (e.g., part-of-speech tagging), which helped to assess how likely they are to describe TRIs. Only sentences containing at least one

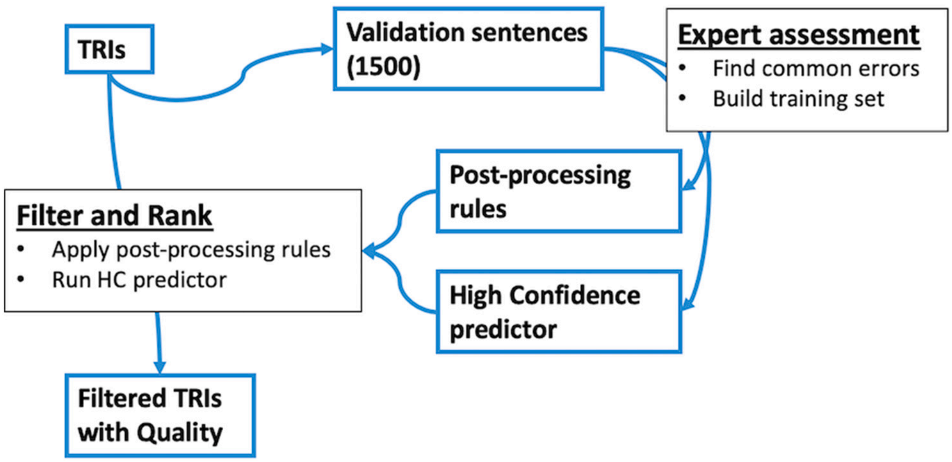

Fig. 10. Overview of post-processing and confidence prediction process. A validation set is formed with 1500 randomly selected sentences, which are manually validated by two independent reviewers. The results of the validation are used to design a set of post-processing rules for the most common errors, as well as to train a high confidence predictor using a random forest model. The original TRIs are filtered using the post-processing rules and labelled with 'High' or 'Low' confidence based on the predictor.

gene/protein mention attributable to a TF from the list of dbTFs as described in TFClass [21,22] were considered further. Selected sentences should also mention at least one other gene, which would correspond to a candidate target gene of that TF. We allowed the target gene to be the same as the TF to capture potential auto-regulation events, although, as we will see later, the majority of those turned out to be false positives. All potential TRI pairs found in these sentences were checked by a classifier and a rule-based system to produce a list of likely TF-TG TRIs.

After the text-mining pipeline was run, extensive post-processing and quality control work was done on the results. In particular, the list of TRIs was subjected to a validation process by two independent expert curators. This validation identified some common errors, which were addressed by some ad-hoc post-processing rules and also allowed us to develop a measure of confidence for each TRI.

4.4. Assessment of confidence of TRIs

The text-mining pipeline produced results that, unsurprisingly, contain a significant number of incorrect TF-TG TRI-entries. To make the results useful in settings that are not forgiving of wrong entries, for instance, the assembly of gene regulation models, we subsequently conducted an effort to provide a better assessment of quality for these TRIs. Working with the results of the text-mining pipeline, we generated a corpus of 1500 validated TRI/sentences, selected randomly from the 100,000 sentences in the complete text-mining result, which we call the *consensus validation dataset* (Supplementary Table 3). Each sentence was validated independently by two experts, essentially following the curation guidelines for the Gold Standard, until consensus was established. While doing this curation, we determined the most common types of mistakes and drafted a protocol that could be automatically applied to the full results set to produce a high confidence subset. After investigating several aspects, such as differences in interaction scores or the coverage of each pair in different PMIDs (see Supplementary Document 5 for more information), this investigation led to a collection of post-processing rules that were used to predict TRI confidence (Fig. 10). Furthermore, as our validation work indicated that only slightly less than 10% of the autoregulation TRIs (cases where the transcription factor is encoded by the target gene) are valid, we flagged all cases of autoregulation to enable users to filter out these TRIs when needing a high confidence TF-TG gene regulation set. In sum, our procedures provide users of ExTRI the opportunity to focus on high confidence TRIs.

4.5. Post-processing rules

Part of the evaluation of the consensus validation dataset was to understand the origin of false positives, i.e., non-TRI sentences included in the results. The most recurrent mistakes had to do with incorrect detection and normalization of mentions to TF or TG. We compiled a list of the most common issues leading to misclassification of a TF-TG pair, which was subsequently used to develop a set of post-processing rules (Supplementary Table 6). Applying these post-processing rules removed almost 8000 entries from the results (the sentences containing removed entries are also provided in Supplementary Table 6).

4.6. Random forest prediction of confidence

Once we completed the work on the consensus validation dataset and the post-processing rules, we applied a Random Forest model trained over the validation dataset and used it to assign an alternative measure of confidence, since the score produced by the maximum entropy classifier (interaction score, see Section 2.5) was insufficient by itself to determine high confidence sentences. This model used features extracted from the ExTRI results deemed relevant during our preliminary evaluation of the pipeline results (see Supplementary Document 5), which we used to define the validation set. The final list of features used for confidence assessment includes interaction score, number of PMIDs containing that same TRI, number of sentences containing that TRI in the same abstract, and number of TRIs present in that sentence. The presence of several abstracts mentioning the TRI, and in particular, the occurrence of the same TRI several times in the same abstract, were good indicators of high confidence, while having a large number of potential TRIs coming from the same sentence was an indicator of low confidence. We evaluated this model using 5-fold cross-validation, achieving an average F1-score of around 0.75. Specifically, 25 repetitions of this cross-validation assessment reveal that, on average, sentences from the manually curated consensus validation dataset classified as high confidence were found to be correct about 75% of the time (Supplementary Table 7). Since this dataset is a random selection representative of the entire corpus and thoroughly curated by experts, this level of correctness can be extrapolated to all high confidence sentences in the ExTRI.

Supplementary data are available at the journal website online and at www.extri.org. Supplementary data to this article can be found online at https://doi.org/10.1016/j.bbagrm.2021.194778.

CRediT authorship contribution statement

Miguel Vazquez: Conceptualization, Methodology, Software, Formal analysis, Writing - Original Draft, Visualization.

Martin Krallinger: Writing - Original Draft.

Florian Leitner: Conceptualization, Methodology, Software, Formal analysis.

Martin Kuiper: Writing - Original Draft, Visualization, Investigation, Dissemination, Validation.

Alfonso Valencia: Conceptualization, Resources, Supervision, Project administration, Writing - Review & Editing.

Astrid Laegreid: Conceptualization, Validation, Investigation, Supervision, Project administration, Writing - Original Draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank the participants of the COST Action GREEKC (CA15205) for fruitful discussions during workshops supported by COST (European Cooperation in Science and Technology).

References

[1] L. Everett, M. Hansen, S. Hannenhalli, Regulating the regulators: modulators of transcription factor activity, Methods Mol. Biol. 674 (2010) 297–312.
[2] M. Levine, R. Tjian, Transcription regulation and animal diversity, Nature 424 (2003) 147–151.
[3] D. Howe, M. Costanzo, P. Fey, et al., Big data: the future of biocuration, Nature 455 (2008) 47–50.
[4] W.A. Baumgartner, K.B. Cohen, L.M. Fox, et al., Manual curation is not sufficient for annotation of genomic databases, Bioinformatics 23 (2007) i41–i48.
[5] Z. Lu, L. Hirschman, Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II, Database (Oxford) 2012 (2012).
[6] A. Cn, R. Pm, A. S, BioCreative III interactive task: an overview, BMC Bioinformatics 12 (Suppl. 8) (2011). S4–S4.
[7] N. Papanikolaou, G.A. Pavlopoulos, T. Theodosiou, et al., Protein–protein interaction predictions using text mining methods, Methods 74 (2015) 47–53.
[8] À. Bravo, J. Piñero, N. Queralt-Rosinach, et al., Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research, BMC Bioinformatics 16 (2015) 55.
[9] J. Li, Y. Sun, R.J. Johnson, BioCreative V CDR task corpus: a resource for chemical disease relation extraction, Database (Oxford) 2016 (2016).
[10] M. Krallinger O. Rabal S.A. Akhondi , et al. Overview of the BioCreative VI Chemical-protein Interaction Track. 6.
[11] J. Saric, L.J. Jensen, P. Bork, Extracting regulatory gene expression networks from pubmed, in: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), Barcelona, Spain, 2004, pp. 191–198.
[12] C. Rodríguez-Penagos, H. Salgado, I. Martínez-Flores, et al., Automatic reconstruction of a bacterial regulatory network using natural language processing, BMC bioinformatics 8 (2007) 1–11.
[13] S. Aerts, M. Haeussler, S. van Vooren, et al., Text-mining assisted regulatory annotation, Genome Biol. 9 (2008), R31.
[14] H.-C. Wang, Y.-H.S. Chen, H.-Y. Kao, Inference of transcriptional regulatory network by bootstrapping patterns, Bioinformatics 27 (2011) 1422–1428.
[15] Y.-T. Tang, S.-J. Li, H.-Y. Kao, et al., Using unsupervised patterns to extract gene regulation relationships for network construction, PLoS ONE 6 (2011), e19633.
[16] F. Leitner, M. Krallinger, S. Tripathi, et al., Mining cis-regulatory transcription networks from literature, Proc. BioLINK SIG 2013 (2013) 5–12.
[17] 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 (2012) S1.
[18] S. Farahmand, T. Riley, K. Zarringhalam, ModEx: a text mining system for extracting mode of regulation of transcription factor-gene regulatory interaction, J. Biomed. Inform. 102 (2020), 103353.
[19] H. Han, H. Shim, D. Shin, et al., TRRUST: a reference database of human transcriptional regulatory interactions, Sci. Rep. 5 (2015) 11432.
[20] J. Devlin, M.-W. Chang, K. Lee, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805 [cs], 2019.
[21] E. Wingender, T. Schoeps, M. Haubrock, et al., TFClass: expanding the classification of human transcription factors to their mammalian orthologs, Nucleic Acids Res. 46 (2018) D343–D347.
[22] E. Wingender, T. Schoeps, M. Haubrock, et al., TFClass: a classification of human transcription factors and their rodent orthologs, Nucleic Acids Res. 43 (2015) D97–D102.
[23] S. Holmås, R. Riudavets Puig, M.L. Acencio, et al., The cytoscape BioGateway app: explorative network building from an RDF store, Bioinformatics 36 (2020) 1966–1967.
[24] S. Holmås, R.R. Puig, M.L. Acencio, et al., The cytoscape BioGateway app: explorative network building from the BioGateway triple store, Bioinformatics 36 (6) (2019) 1966–1967.
[25] P. Sassone-Corsi, Coupling gene expression to cAMP signalling: role of CREB and CREM, Int. J. Biochem. Cell Biol. 30 (1998) 27–38.
[26] D. De Cesare, P. Sassone-Corsi, Transcriptional regulation by cyclic AMP-responsive factors, Prog. Nucleic Acid Res. Mol. Biol. 64 (2000) 343–369.
[27] V. Touré S. Vercruysse M.L. Acencio , et al. The minimum information about a molecular interaction causal statement (MI2CAST). Bioinformatics.
[28] A. Essaghir, F. Toffalini, L. Knoops, et al., Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data, Nucleic Acids Res. 38 (2010), e120.
[29] S. Kerrien, B. Aranda, L. Breuza, et al., The IntAct molecular interaction database in 2012, Nucleic Acids Res. 40 (2012) D841–D846.
[30] R.P. Huntley, T. Sawford, P. Mutowo-Meullenet, et al., The GOA database: gene ontology annotation updates for 2015, Nucleic Acids Res. 43 (2015) D1057–D1063.
[31] H. Han, J.-W. Cho, S. Lee, et al., TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions, Nucleic Acids Res. 46 (2018) D380–D386.
[32] L. Perfetto, L. Briganti, A. Calderone, et al., SIGNOR: a database of causal relationships between biological entities, Nucleic Acids Res. 44 (2016) D548–D554.
[33] S. Carrasco Pro, A. Dafonte Imedio, C.S. Santoso, et al., Global landscape of mouse and human cytokine transcriptional regulation, Nucleic Acids Res. 46 (2018) 9321–9337.
[34] T. Huang, X. Huang, B. Shi, et al., GEREDB: gene expression regulation database curated by mining abstracts from literature, J. Bioinforma. Comput. Biol. 17 (2019), 1950024.