Computational and Structural Biotechnology Journal 20 (2022) 4636–4644

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

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

Method article

YTLR: Extracting yeast transcription factor-gene associations from the literature using automated literature readers

Tzu-Hsien Yang [a],[*], Chung-Yu Wang [b],[1], Hsiu-Chun Tsai [b],[1], Ya-Chiao Yang [b],[1], Cheng-Tse Liu [b]

[a] *Department of Biomedical Engineering, National Cheng Kung University, No.1, University Road, Tainan 701, Taiwan*
[b] *Department of Information Management, National University of Kaohsiung, Kaohsiung University Rd, 811 Kaohsiung, Taiwan*

ARTICLE INFO

Article history:
Received 6 May 2022
Received in revised form 16 August 2022
Accepted 18 August 2022
Available online 24 August 2022

Keywords:
Transcriptional regulation
BioBERT
Natural language processing

ABSTRACT

Cells adapt to environmental stresses mainly via transcription reprogramming. Correct transcription control is mediated by the interactions between transcription factors (TF) and their target genes. These TF-gene associations can be probed by chromatin immunoprecipitation techniques and knockout experiments, revealing TF binding (TFB) and regulatory (TFR) evidence, respectively. Nevertheless, most evidence is still fragmentary in the literature and requires tremendous human resources to curate. We developed the first pipeline called YTLR (Yeast Transcription-regulation Literature Reader) to automate TF-gene relation extraction from the literature. YTLR first identifies articles with TFB and TFR information. Then TF-gene binding pairs are extracted from the TFB articles, and TF-gene regulatory associations are recognized from the TFR papers. On gathered test sets, YTLR achieves an AUC value of 98.8% in identifying articles with TFB evidence and AUC = 83.4% in extracting the detailed TF-gene binding pairs. And similarly, YTLR also obtains an AUC value of 98.2% in identifying TFR articles and AUC = 80.4% in extracting the detailed TF-gene regulatory associations. Furthermore, YTLR outperforms previous methods in both tasks. To facilitate researchers in extracting TF-gene transcriptional relations from large-scale queried articles, an automated and easy-to-use software tool based on the YTLR pipeline is constructed. In summary, YTLR aims to provide easier literature pre-screening for curators and help researchers gather yeast TF-gene transcriptional relation conclusions from articles in a high-throughput fashion. The YTLR pipeline software tool can be downloaded at https://github.com/cobisLab/YTLR/.

1. Introduction

Transcription reprogramming controls cellular fitness when environmental changes happen [[1,2,3]. Under stress conditions, the expression of different sets of genes helps cells respond to external stimuli [4,5]. The precise regulation of gene transcription is usually mediated by the binding of transcription factors (TFs) to the promoter regions of their target genes [6,7]. Malfunction of TFs can lead to abnormal cellular traits or even cell death. Hence, understanding the correct TF-gene relations is a fundamental goal in molecular biology.

There are two experimental approaches to understanding the target genes of TFs [8,9]. First, researchers can probe the binding target sequences of specific TFs using chromatin immunoprecipitation (ChIP). These TF-bound sequences are then mapped to the promoter regions, the proximal genic regions, or the distal regulatory sites of their target genes [6,10]. Articles containing these TF binding evidence are called TF binding (TFB) evidence literature. Second, the affected target genes of a given TF can be identified from the expression changes between the controlled wild-type samples and the knocked-out/depleted/altered/enhanced samples of this particular TF [2,11–13]. In this case, the identified genes are indirectly regulated by the given TF through certain mechanisms [8]. Works describing the TF indirect regulatory gene target information are called TF regulation (TFR) evidence literature. And it has been shown that regulation mechanism hypotheses can be inferred via integrating the TF-gene binding pairs from TFB articles and the TF-gene regulatory associations from TFR papers [7,8,14].

However, most of these experimentally verified TF-gene binding or regulatory association knowledge is fragmentary in the literature. The lack of an automated way to collect these TF-gene relations hinders a comprehensive understanding of transcription

[*] Corresponding author.

E-mail addresses: thyangza@gs.ncku.edu.tw (T.-H. Yang), a1073333@mail.nuk.edu.tw (C.-Y. Wang), a1073316@mail.nuk.edu.tw (H.-C. Tsai), a1073314@mail.nuk.edu.tw (Y.-C. Yang), a1073311@mail.nuk.edu.tw (C.-T. Liu).

[1] These authors contributed equally.

https://doi.org/10.1016/j.csbj.2022.08.041
2001-0370/© 2022 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.

regulation [15]. YEASTRACT (Yeast Search for Transcriptional Regulators And Consensus Tracking) is the first database that deposits more than 175,000 manually curated yeast TF-gene associations [9]. They have united the efforts of many researchers in reading the yeast-related articles and sorted out valuable TF-gene binding or regulatory evidence. While manual literature curation can collect some TFB and TFR evidence, rapid progress in the community overwhelms the available human resources. For example, over 2,400 yeast-related research articles are deposited in the PubMed database annually. This rapid research progression makes automated literature machine readers suitable for lowering the burden on human curators [16]. Therefore, there is a need to build auto-literature machine readers for organizing and gathering the most recent experimental results related to transcriptional regulation from the literature.

Researchers have developed event extraction and name entity recognition algorithms to help understand human written languages. These tools identify the possible noun and verb phrases in articles and link the potentially related entities together [17–21]. Unfortunately, general sentence event extraction methods cannot specifically identify TF-gene transcriptional associations, leading to unacceptable accuracy in grasping TF-gene transcriptional regulation research articles. On the other hand, literature name-entity machine annotators can only mark the existence of TFs and genes. They provide no TF-gene association information at all. Currently, there is no appropriate auto-literature reader pipeline for extracting TF-gene binding or regulatory association conclusions from the literature.

In this research, we developed a deep learning pipeline called YTLR (Yeast Transcription-regulation Literature Reader) to automate TF-gene association conclusion extraction from the literature. YTLR comprises in–house software tools and two-phased deep learning literature readers to recognize TF-gene binding and regulatory associations. Articles describing TFB and/or TFR evidence are first identified by the deep literature identification networks in YTLR Phase I. For the articles identified to have TFB evidence, TF-gene binding pairs in these articles are recognized in YTLR Phase II. And TF-gene regulatory associations in papers classified to describe TFR results are also summarized in this phase. On the set-aside test sets, we demonstrated that the auto-literature readers achieve high performance in identifying articles with TFB/TFR evidence (YTLR Phase I AUC = 98.8%/98.2%) and in recognizing TF-gene binding/regulatory associations (YTLR Phase II AUC = 83.4%/80.4%). Compared with related works, YTLR Phase I auto-literature readers outperform previously proposed models by at least 4.3%/2.4% AUC values in identifying TFB/TFR articles. And YTLR Phase II deep networks obtain at least 26%/30% better AUC values than existing baseline methods in recognizing TF-gene binding/regulatory associations. To facilitate researchers in extracting TF-gene transcriptional relations from large-scale queried articles, an automated and easy-to-use software tool is constructed based on the YTLR pipeline. Lastly, we reported that the miss rate of YTLR for identifying TFB and TFR articles is estimated to be only at most around 9% larger than human curators while saving a tremendous number of human resources. In summary, YTLR is a tool that facilitates easier literature pre-screening for curators and helps researchers collect yeast TF-gene transcriptional relation conclusions from articles in a high-throughput fashion. The YTLR software tool can be freely downloaded at https://github.com/cobisLab/YTLR/.

2. Methods and Datasets

We constructed the automated literature-reader pipeline named YTLR (Yeast Transcription-regulation Literature Reader) based on the BioBERT (Bidirectional Encoder Representations from Transformers) pre-trained model and the transfer learning technique using the ground truth datasets collected from YEASTRACT. The workflow of YTLR can be divided into three steps (See Fig. 1): (I) Pre-process the given files. (II) YTLR Phase I: Identify articles with TFB and/or TFR evidence. Based on the abstract of the given article, it is first tagged as a TFB article if it contains TFB evidence and classified as a TFR paper if it describes TFR results. (III) YTLR Phase II: Recognize TF-gene binding pairs and regulatory associations. We first extract the sentence descriptions for all TF-gene pairs found in the full text of a TFB article. These sentence descriptions are then classified if they contain TF-gene binding information. Finally, based on the classification summaries of these sentence descriptions, the extracted TF-gene pairs from TFB articles are summarized as binding relations or random pairs. A similar process is applied to TFR articles to obtain TF-gene regulatory associations. Details of data preparation and YTLR steps are elucidated in the following sections.

2.1. Collection of the ground-truth datasets

YEASTRACT has manually curated the TFB and TFR literature evidence for TF-gene transcriptional associations in *Saccharomyces cerevisiae*. We downloaded the TF-gene binding pairs with their corresponding TFB articles and the TF-gene regulatory associations with the corresponding TFR articles from YEASTRACT (2021 Feb repository). Three datasets were prepared from the YEASTRACT data: (1) the articles curated to have TFB evidence and papers describing TFR results; (2) the TF-gene sentence descriptions; and (3) the exact TF-gene binding pairs from TFB articles and TF-gene regulatory associations from TFR articles.

2.1.1. The TFB article and TFR paper datasets

The TFB article and TFR paper datasets were generated based on the downloaded transcription regulation-related literature curated by YEASTRACT. When a TF-gene binding pair is curated from an article, this article is tagged to contain TFB evidence. And when a TF-gene regulatory pair is curated from an article, we tagged this article as a TFR article. In total, we gathered 463 articles curated to have TFB evidence and 1,196 papers curated to convey TFR evidence in this research. The abstracts of these TFB articles and TFR papers form the positive sets of the TFB and TFR article datasets. To obtain the negative set, we first collected articles published before 2018 with titles containing the word ”yeast” or ”*Saccharomyces cerevisiae*” from PubMed. The articles tagged with TFB or TFR evidence were then eliminated from this negative set. 72,502 yeast-related articles that were not curated to have TFB or TFR evidence were organized as the negative set. We partitioned these articles into TFB and TFR training-validation sets and test sets based on their publication years. The collected articles published before 2014 were used to form the model training-validation ground-truth dataset. The TFB article training-validation set includes 379 articles with TFB evidence as the TFB positive samples, and the TFR article training-validation set encompasses 970 papers with TFR evidence as the TFR positive samples. 58,135 papers published before 2014 with no curated TFB or TFR evidence were gathered as the overall negative samples. And papers published between 2014 and 2018 were cut out as the test sets for performance generalization evaluation. The TFB test set includes 84 articles with TFB evidence as TFB positive test samples, and the TFR test set has 226 papers with TFR evidence as TFR positive test samples. To prepare the negative samples for the TFB and TFR test sets, we sampled 84 and 226 articles from the 14,453 yeast-related papers published during 2014–2018 that have no curated TFB or TFR evidence as the TFB and TFR negative test samples, respec-

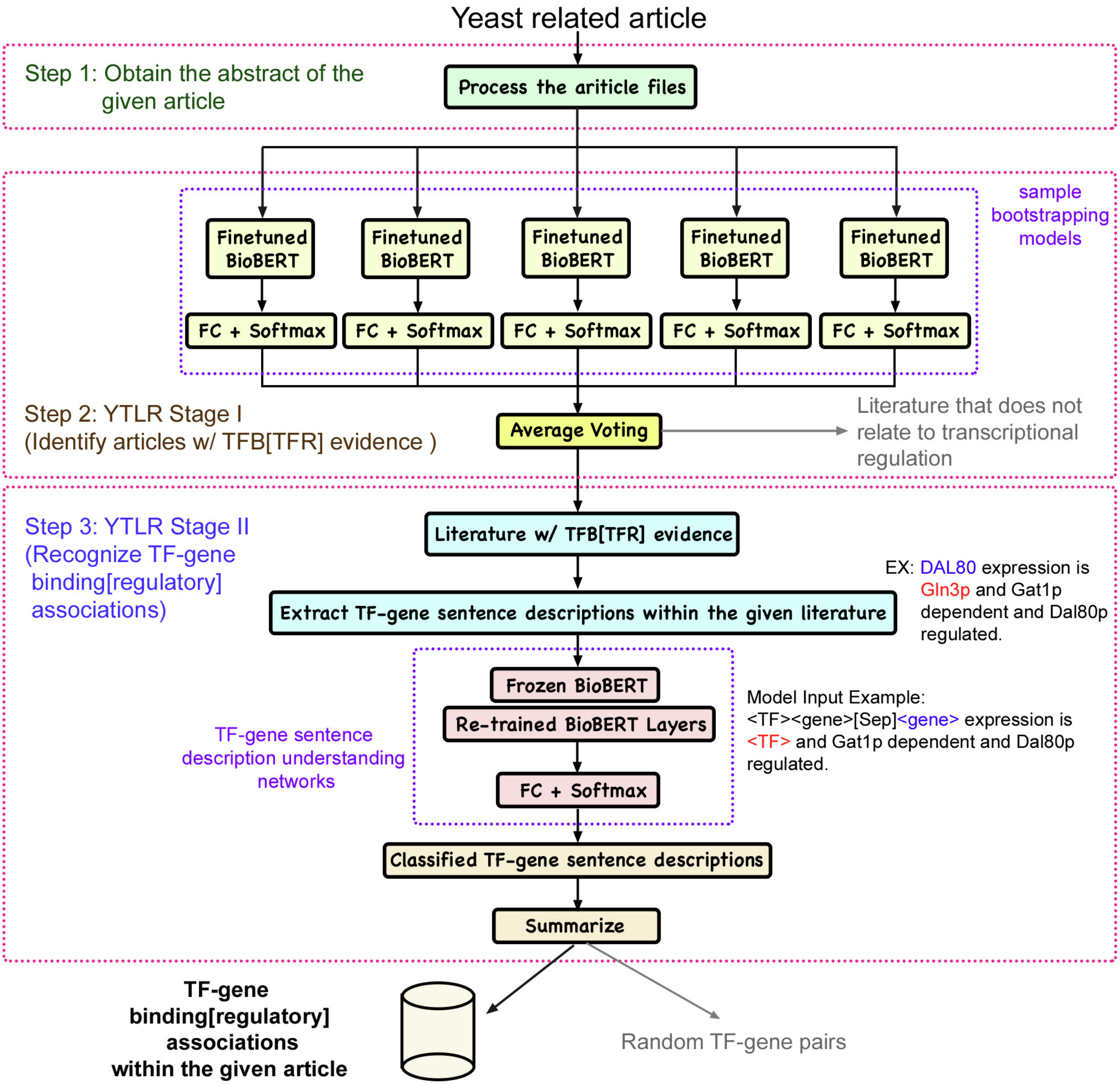
Yeast related article
Step 1: Obtain the abstract of the given article
Process the ariticle files
sample bootstrapping models
Finetuned BioBERT
FC + Softmax
Step 2: YTLR Stage I (Identify articles w/ TFB[TFR] evidence)
Average Voting
Literature that does not relate to transcriptional regulation
Step 3: YTLR Stage II (Recognize TF-gene binding[regulatory] associations)
Literature w/ TFB[TFR] evidence
EX: DAL80 expression is Gln3p and Gat1p dependent and Dal80p regulated.
Extract TF-gene sentence descriptions within the given literature
Frozen BioBERT
Re-trained BioBERT Layers
FC + Softmax
TF-gene sentence description understanding networks
Model Input Example: <TF><gene>[Sep]<gene> expression is <TF> and Gat1p dependent and Dal80p regulated.
Classified TF-gene sentence descriptions
Summarize
TF-gene binding[regulatory] associations within the given article
Random TF-gene pairs

Fig. 1. The overview of the YTLR pipeline in identifying TF-gene binding pairs and TF-gene regulatory associations. Similar architectures were built and performed separately in YTLR for these two types of transcriptional TF-gene relations (TF-gene binding relations from the TFB literature and TF-gene regulatory associations from the TFR literature). These two independent pipelines were denoted by brackets (i.e., pipelines for extracting TF-gene binding[regulatory] pairs from the TFB[TFR] articles).

tively. In YTLR Phase I, only the abstracts of the articles are utilized for identifying literature with TFB and TFR evidence.

2.1.2. The TF-gene sentence description datasets

TF-gene sentence description datasets were prepared based on the YEASTRACT-curated TF-gene binding pairs and TF-gene regulatory associations. A TF-gene sentence description in the PMC-retrieved full text of a TFB article or a TFR paper consists of one or two consecutive sentences that contain the given TF and gene. The lists of yeast genes and TFs were collected from the SGD database [22], the work of Harbison *et al.* [6], and YEASTRACT [9]. Sentences in the ”Materials and Methods” section, the ”Supplementary File” section, and the ”Reference” section are excluded to avoid external information inclusion. A TF-gene sentence description from a TFB article is denoted to contain binding information if this TF-gene pair was curated by YEATRACT to be a binding pair in the article. Otherwise, the TF-gene sentence description is annotated not to contain binding information. Similar procedures were applied to obtain the TF-gene regulatory sentence descriptions from TFR papers. We picked the TF-gene sentence descriptions from articles published before 2014 as the training-validation ground-truth datasets. In summary, 11,872 and 24,571 sentence descriptions for TF-gene binding pairs and TF-gene regulatory associations were included in the training-validation ground-truth datasets, respectively. And we sampled 15,113 sentence descriptions of TF-gene non-binding pairs and 27,411 sentence descriptions of TF-gene non-regulatory pairs from the same articles. These sentence description datasets were used to train the two sentence-description understanding networks that recognize TF-gene binding pairs and regulatory associations.

2.1.3. The TF-gene binding pair and regulatory association test sets

The overall YTLR Stage II recognizes the TF-gene binding pairs and regulatory associations from the TFB articles and TFR papers, respectively. To help evaluate the recognition performance of YTLR Stage II, TF-gene pairs found in the papers published during 2014–2018 were separated to serve as the TF-gene association test set. TF-gene transcriptional associations were adopted from the YEASTRACT manual curation results. In total, 472 TF-gene binding pairs along with 472 sampled non-binding pairs found in the same TFB articles were separated to form the TF-gene binding pair test set. And 735 regulatory associations along with 731 sampled non-regulatory TF-gene pairings found in the same TFR articles were separated to form the TF-gene regulatory association test set.

2.2. The YTLR TF-gene association extraction pipeline

YTLR can identify TF-gene binding pairs and regulatory associations based on two networks built on similar architectures (See Fig. 1). The TF-gene binding and regulatory relations are extracted from the TFB articles and TFR papers, respectively. There are two phases of auto literature readers in YTLR. Take the TF-gene binding pair extraction workflow as an example. For a given article, YTLR first checks if it provides TFB evidence based on its abstract (YTLR Stage I). Then TF-gene pairs and the related sentence descriptions are extracted from the articles identified to have TFB evidence using a python script. The extracted TF-gene sentence descriptions from TFB articles are checked to see whether they portray TF-gene binding information. At the end of YTLR Stage II, the TF-gene binding pairs or random pairings for the given TFB articles are summarized based on the classified sentence descriptions. Similar procedures are also performed on TFR articles to obtain the TF-gene regulatory pairs in YTLR.

2.2.1. Phase I: identification of literature with TFB/TFR evidence

Two deep networks were constructed to identify TFB articles and TFR papers. We articulate the construction of the TFB article identification network here as an example. Because of the severely imbalanced positive and negative sample numbers in the TFB article datasets, we designed a sample-bootstrapping ensemble deep learning architecture for obtaining a high-performance TFB literature identification network. Based on the abstracts of the given articles, TFB deep network annotates if the provided literature contains TFB evidence. The designed sample-bootstrapping ensemble deep learning architecture is sketched in Fig. 1-Step 2. The TFB auto tagger was trained using the TFB article training-validation set formed by the 379 articles with TFB evidence and the 58,135 articles that have neither curated TFB nor TFR evidence. Since the cardinality of the negative set out-numbers that of the positive set, we applied the sample bootstrapping method to down-sample the negative set and obtained an ensemble model for overcoming the imbalance problem [23,24]. We repeated the bootstrap-sampling five times to obtain five diverse negative subsets that all have the same cardinality as the positive set. The subsets were sampled to enforce a similar year distribution to the positive set in the bootstrapping sampling process. Then based on the positive set and the five negative subsets, 5 sample-bootstrapping TFB deep models were constructed to provide probabilities that the given articles contain TFB evidence, respectively. In each sample-bootstrapping deep model, the BioBERT [25] network was adopted as the building block. The *i*th bootstrapping model based on the *i*th bootstrapping dataset can be written as the following:

$$p_i = \mathrm{softmax}(C_i W_i),\ C_i = \mathrm{FineTuned\_BioBERT_i}(A),$$

where A is the word piece tensor for the abstract of the given article, C_i is the summarizing CLS (classifier token) vector generated by the *i*th fine-tuned BioBERT model, W_i is the trainable weight matrix for the *i*th sample-bootstrapping model, and p_i is the TFB tagging probability given by the *i*th sample-bootstrapping model. We initialized this building block using the pre-trained BioBERT weights. And the whole weights were fine-tuned on the positive set and one bootstrap-downsampled negative subset. Fivefold cross-validation on the positive set and each bootstrap-sampled negative subset was used for selecting the hyperparameters of each sample-bootstrapping model. Finally, these five probabilities were averaged to aggregate the final TFB identification results. The sample-bootstrapping ensemble deep networks aim to provide robust TFB literature identifiers among diverse biomedical papers. The same procedure was also applied to the 970 papers with TFR evidence in the TFR article training-validation set to build a TFR literature identification network.

2.2.2. Phase II: recognition of TF-gene binding/regulatory associations

After identifying articles with TFB evidence based on their abstracts, we further try to recognize the precise TF-gene binding associations from the article. A similar process is also done for TFR articles to extract TF-gene regulatory pairs. We explain the method for recognizing TF binding pairs as an example in this subsection.

YTLR Phase II (See Fig. 2-Step3) first extracts the sentences that potentially describe the association between a TF and a gene from the articles identified to have TFB evidence. We assume that the authors describe the annotated TF-gene binding associations with a distinguishable tone from random pairings. Hence the extracted sentence descriptions are then fed into the deep TF-gene sentence description understanding network to check if they contain binding information. Finally, based on the classified sentence descriptions, the TF-gene pairs are recognized as binding associations or random pairings. We used the gathered TF-gene sentence description training-validation sets to train the TF-gene sentence description understanding deep networks. Two deep networks were constructed for YTLR Phase II. The TF-gene binding sentence description understanding network was built to check if a TF-gene sentence description extracted from a TFB article contains binding information for a TF-gene pair. Similarly, the TF-gene regulatory sentence description understanding network was constructed to see if a TF-gene sentence description from a TFR article shows regulation information for a TF-gene pair. The TF-gene binding understanding networks are formulated as

$$p = \mathrm{softmax}(FW),\ F = f_1 \circ \mathrm{Frozen\_BioBERT}(S),$$

where S is the word piece tensor for the given TF-gene sentence description, $\circ$ denotes function composition, f_1 is the function formed by the tunable BioBERT last five layers (which includes the attention operation, dropout, layer-normalization, and dense sub-layers), F is the summarizing CLS (classifier token) vector, W is the trainable weight matrix, and p is the probability that the TF-gene sentence description conveys binding information for this TF-gene pair. To fairly divide the sample importance into each TF-gene pair, we enforced the optimization loss of the sentence understanding networks to obey the following specification:

$$L = \frac{1}{N}\sum_{i=1}^{N} w_i L_i, w_i = \frac{1}{k_{j\_i}},$$

where L is the training loss, L_i is the softmax loss for the sentence description i, k_{j_i} is the number of sentence descriptions related to the TF-gene pair j that is in the same article as the sentence description i, and N is the total number of sentence descriptions. When training the sentence description understanding networks, fivefold cross-validation was applied to the sentence description training-validation sets.

The final confident TF-gene binding associations for a given TFB article are summarized using the following equation:

$$TG_j = \{\text{TF\_gene pair}_{ji} \mid \#P_{ji} > \#N_{ji}\},$$

where TG_j represents the set of TF-gene binding associations in the jth TFB article, P_{ji} denotes the collection of sentence descriptions classified to contain binding information for the ith TF-gene pair in the jth TFB article, N_{ji} is the set of sentence descriptions classified to not relate to transcriptional binding for the ith extracted TF-gene pair in the jth TFB article, and #(.) denotes the cardinality of the given set. After YTLR Phase II, TF-gene binding pairs are discriminated from random TF-gene pairings for a given TFB article. The same procedure was also utilized to extract TF-gene regulatory pairs from the identified TFR articles.

2.3. Hyperparameters of YTLR

Hyperparameters of YTLR were chosen by fivefold cross-validation. We adopted the following hyperparameters in fine-tuning the two deep identification networks for TFB and TFR articles in YTLR Phase I: (1) learning rate schedule: step linear warm-up followed by cosine decay (max learning rate = 1e-5); (2) optimization method: Adam; (3) number of epochs: 4; (4) neuron initialization: pre-trained BioBERT; (5) batch training size: 16. Dropout layers (dropout rate = 0.1) were added to regularize the training process and prevent over-fitting that may kill model generalization. And hyperparameters used in training the two deep networks for understanding TF-gene binding and regulatory sentence descriptions in YTLR Phase II were as follows: (1) learning rate schedule: 1e-5 followed by cosine decay; (2) optimization method: Adam; (3) number of epochs: 10; (4) neuron initialization: Phase I fine-tuned BioBERT; (5) batch training size: 16; (6) dropout: 0.2 for both the hidden state layers and attention layers, and 0.5 for the last classification layer. The overall YTLR pipeline was trained using NVIDIA RTX Titan GPUs.

3. Results and Discussions

3.1. Performance of YTLR

The constructed automated literature readers in YTLR were implemented through two phases to extract TF-gene binding and regulatory associations. First, the specified article is checked to see if it contains TFB and/or TFR evidence based on its abstract. For those articles identified to have TFB evidence, all TF-gene pairs are extracted from the PMC-retrieved full texts. These TF-gene pairs are then recognized to be binding pairs or random pairings based on the extracted TF-gene sentence description summaries. Similar procedures are applied to TFR articles for extracting TF-gene regulatory associations. We utilized the following metrics for estimating the performance of YTLR Phase I (TFB and TFR article identification) and Phase II (TF-gene binding and regulatory association recognition) [26,27]:

$$\text{Recall(Sensitivity)} = \frac{TP}{TP + FN},$$

$$\text{Specificity} = \frac{TN}{FP + TN} = 1 - \text{FPR},$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{F1} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}},$$

where FPR stands for the false-positive rate, and TP, FP, TN, and FN abbreviate for true positive, false positive, true negative, and false negative calculated under a specified threshold on the prediction probability, respectively. TP/TN count the numbers of correctly identified samples from the positive/negative sets, and FP/FN sum up the numbers of mistakenly annotated samples from the negative/positive sets. The F1 value is the harmonic mean of the calculated precision and recall values. It can help evaluate the overall balance between the precision and recall trade-off under a specific threshold. Since the recall, specificity, precision, and F1 values are threshold-specific, in this research, we selected a general prediction probability threshold of 0.5 in calculating these metrics. To further eliminate the threshold effect when estimating the intrinsic capability of a prediction model, we also considered receiver operating characteristic (ROC) curves. The ROC curve plots Sensitivity values (i.e., Recall) against the corresponding (1 - Specificity) values (i.e., FPR) when the prediction thresholds are varied. In this sense, the ROC curve evaluates the intrinsic capability of a model under different chosen thresholds. Because of the nature of ROC curves, they can be used for fair model comparisons that eliminate the threshold effects of different prediction methods. The more upper-left the ROC curve is observed, the better discrimination power the model can achieve for the specified task, meaning that the model achieves a high recall even if the threshold is controlled to have a low FPR. This result can be estimated by calculating the area under the ROC curve (AUC). We report these results parallelly in the following subsections since two networks were built and executed independently to recognize TF-gene binding pairs and regulatory associations from the TFB articles and TFR papers, respectively.

3.1.1. Performance of YTLR Phase I

In YTLR Phase I, a given yeast-related article is checked to see if it describes (1) TFB and/or (2) TFR evidence based on its abstract. In the training-validation process of YTLR Phase I, we fine-tuned the networks built from the pre-trained BioBERT for only four epochs to avoid model over-fitting when optimizing the full BioBERT model. To evaluate the performance and generalization of the TFB literature identification network, we used the test sets formed by the articles published during 2014–2018 in the TFB article dataset. The TFB article test set has 84 articles with curated TFB evidence as positive samples. And the same numbers of sampled non-transcription-related publications are included in the TFB article test set as negative samples. The test performance metric summary for the TFB sample-bootstrapping averaging model is listed in Table 1. As shown in Table 1, the final aggregation result of identifying articles with TFB evidence achieves an AUC value of 98.8% and an F1 value of 94.4% on the test set. The high test AUC value means that YTLR Phase I is well generalized on newly collected articles and can obtain a high TFB literature identification power while the false discovery rate is controlled. Similarly, to evaluate the TFR literature identification network, 226 TFR papers and the same number of non-transcription-related papers published during 2014–2018 were included in the TFR article test set. Similar conclusions can also be made for identifying articles with TFR evidence in YTLR Phase I (AUC = 98.2% and F1 = 92.4%, also see Table 1). Summarizing these test performance results, we can see that users can confidently identify if the given article contains TFB and/or TFR evidence based on its abstract via YTLR Phase I.

3.1.2. Performance of YTLR Phase II

For an identified TFB article, YTLR Phase II first extracts the sentence descriptions for all TF-gene pairs within its full text. Then the TF-gene sentence descriptions are automatically reasoned by deep understanding networks to recognize if they are related to TF-gene binding relations. Finally, these results are summarized to check if the TF-gene pairs are binding associations or random pairings. Similar procedures were also applied to extract TF-gene regulatory associations from the identified TFR papers. Learning curves were used in the fivefold cross-validation process to ensure convergence

Table 1
The performance summary of YTLR Phase I in identifying articles with TFB and TFR evidence on the test sets.

Phase I: evidence literature identification	AUC	F1	Precision	Recall	Specificity
TFB articles	98.8%	94.4%	98.7%	90.5%	98.8%
TFR articles	98.2%	92.4%	93.2%	91.6%	93.4%

and well-fitting of the two deep networks that recognize sentences describing TF-gene binding pairs and regulatory associations. In Figure S1-a and S1-b of the Supplementary File, the training and validation learning curves show that the two deep networks are well-fitted and converge to the optimal configurations. To evaluate the correctness of the final extracted and summarized TF-gene binding and regulatory pairs, we resorted to ROC curves of the results. YTLR Phase II was evaluated on the reserved TF-gene association test sets (Table 2). The test AUC values for extracting and summarizing the exact TF-gene binding pairs and regulatory associations are 83.4% and 80.4%, respectively. Other performance metrics can be found in Table 2. These results show that YTLR Phase II can recognize TF-gene binding and regulatory associations from random pairings with good discrimination power.

3.2. Comparison with previous models and baseline methods

YTLR is the first complete pipeline for automated extraction of yeast TF-gene binding and regulatory associations from the literature using machine readers. In YTLR Phase I, YTLR identifies if the given yeast-related article provides TFB evidence and TFR results based on its abstract. And in YTLR Phase II, TF-gene binding associations are recognized from an article if this article is identified to have TFB evidence. And TF-gene regulatory associations are also recognized from the TFR articles. While no software can carry out the same goals as YTLR, some previous models and baseline methods were designed to provide functions close to parts of YTLR. We compared these models and methods with the corresponding parts of YTLR.

3.2.1. Comparison among YTLR Phase I and similar models

We first compared YTLR Phase I with related models in identifying literature with TFB and TFR evidence. In the work of Burns *et al.* [28], they proposed models that combine word embeddings with convolution neural networks (CNNs) or long short-term memory (LSTM) networks to unravel the existence of molecular interaction descriptions in the literature. To provide a fair comparison among YTLR Phase I models, the CNN models, and the LSTM networks, we retrained the CNN and LSTM models designed by Burns *et al.* using the BioBERT word embeddings and the model architectures designed by the authors. Learning curves for the CNN models and LSTM networks were monitored to ensure model well-fitting and convergence. Details of the learning curves and hyperparameters of the CNN models and LSTM networks can be found in Figure S2 and Figure S3 of the Supplementary File. On the TFB article test set, YTLR TFB literature identification results (AUC = 98.8%) outperform the corresponding CNN classifier by 4.3% (98.8–94.5%) and the corresponding LSTM network by 13.9% (98.8–84.9%) in AUC values (see Fig. 2-a). And on the TFR article test set, YTLR TFR literature identification results (AUC = 98.2%) also achieve 2.4% (98.2–95.8%) and 9.9% (98.2–88.3%) AUC performance boost over the CNN model and the LSTM network (see Fig. 2-b). These results demonstrated that YTLR Phase I provides better transcription-related literature identification performance than previously proposed models.

3.2.2. Comparison among YTLR Phase II and baseline methods

We next compared YTLR Phase II with one baseline method and one language model. The REACH (Reading and Assembling Contextual and Holistic Mechanisms from Text) language model system [29] is selected since it is currently the only available association extraction tool that can handle TF-gene binding pair and regulatory association recognition. We also implemented the simple occurrence-counting metrics of TF-gene pairs as a baseline method. The counting metric is computed by summing up the occurrence of the given TF-gene pair that appears within two consecutive sentences in the full text of a given article. The ROC curves of REACH, the occurrence counting baseline method, and YTLR Stage II results on the TF-gene binding pair and TF-gene regulatory association test sets are summarized in Fig. 3. As shown in Fig. 3-a, YTLR Stage II obtains better AUC values in recognizing TF-gene binding pairs (83.4%) than both the full-text TF-gene occurrence counting metric baseline (57.11%) and the REACH system (55.5%) on the TF-gene binding pair test set. And for TF-gene regulatory association extraction, YTLR Phase II (80.4%) also outperforms the full-text TF-gene occurrence-counting metric baseline (48.77%) and the REACH system (50.3%) on the TF-gene regulatory association test set (See Fig. 3-b). In summary, YTLR Stage II is the leading tool to recognize TF-gene binding pairs and regulatory associations from the literature and outperforms existing tools in grasping the detailed TF-gene transcriptional associations.

3.3. The YTLR software tool facilitates the extraction of TF-gene transcriptional association conclusions from large-scale articles

YTLR is designed to serve as an automated machine reader that extracts the potential TF-gene transcriptional relations from massive numbers of queried articles. For given queried papers, the constructed pipeline helps users narrow down the possible TF-gene binding pairs and regulatory associations for the articles that deserve further detailed investigation. Therefore, YTLR serves to help perform the transcriptional literature pre-screening that facilitates downstream curation or customized knowledge base construction. The tool does not aim to replace or surpass human curation. It is more of a helper tool to facilitate and speed up the curation process.

To facilitate researchers in applying the constructed YTLR pipeline to large-scale yeast article information extraction, we implemented the YTLR software tool with two extra features in addition to the auto-machine literature reader pipeline to lessen the burden of data preparation: (1) Collected summary of TF-gene pairs from large-scale input articles; (2) Automated retrieval of the major measurement methods and experimental conditions. First of all, YTLR aims to provide automated TF-gene transcriptional relation machine readers that can deal with massive numbers of articles. Therefore, we implemented the batch processing function for large-scale article information retrieval. For YTLR Stage I, users can utilize the PubMed function to download all abstracts of the selected (or queried) articles in a `.txt` file for large-scale processing. And YTLR also automates the full-text retrieval process through the PMC-provided FTP services. In YTLR Stage II, users can either rely on our script, which is based on the PMC-provided FTP services, to automatically obtain the available full texts of the provided abstracts or download the `.html` files of

Table 2
The performance summary of YTLR Phase II in recognizing the TF-gene binding pairs and regulatory relations on the test sets.

Phase II: TF-gene association recognition	AUC	F1	Precision	Recall	Specificity
TF-gene binding association	83.4%	81.9%	80.5%	83.3%	79.7%
TF-gene regulatory association	80.4%	80.7%	75.4%	86.8%	71.8%

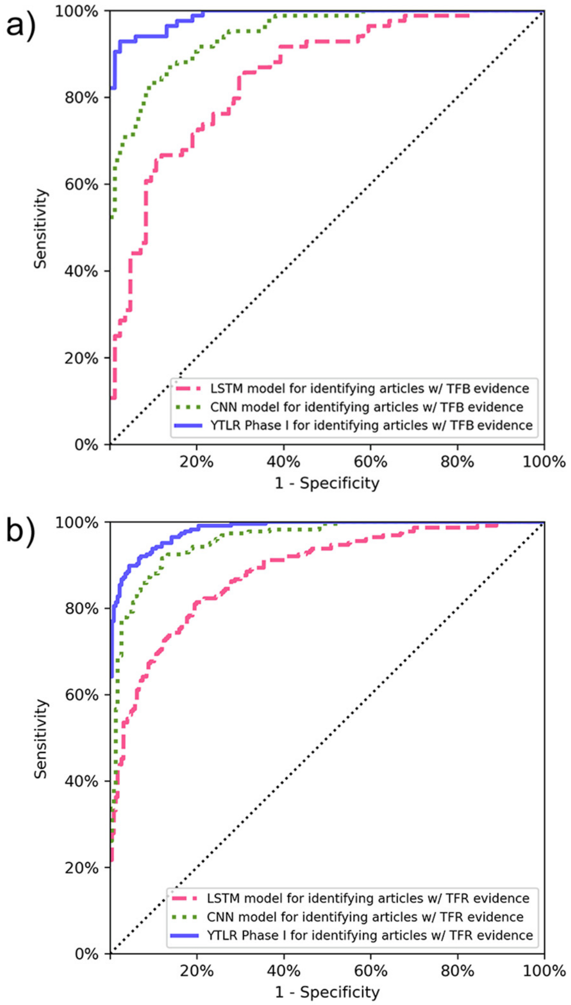

Fig. 2. The test ROC curve comparison among tools that can identify articles with (a) TFB or (b) TFR evidence.

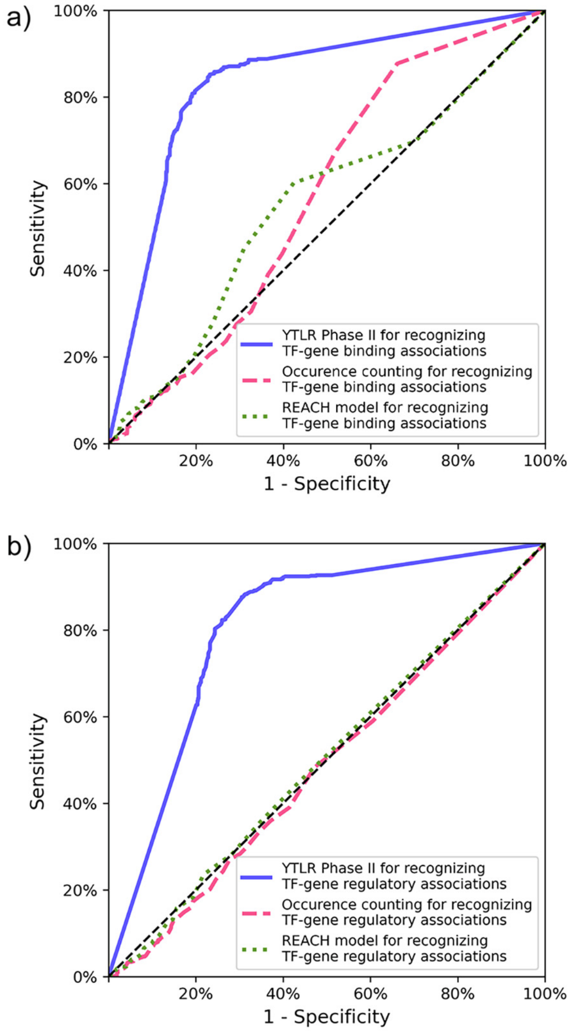

Fig. 3. The test ROC curve comparison among methods in recognizing (a) TF-gene binding pairs from the articles with TFB evidence and (b) TF-gene regulatory pairs from the articles with TFR evidence.

the full texts by themselves. Since the term usage of PMC does not allow web scrawling on their contents, YTLR utilizes only the allowed FTP services in PMC and does not provide website scrawling functions. Based on the available full texts or abstracts, YTLR Stage II outputs the summary of the extracted TF-gene binding pairs and regulatory associations from the massive numbers of TFB and TFR articles for users. The summary of extracted TF-gene pairs can help users understand the involving TFs and genes for the queried cellular situations of the provided article pools.

The second additional feature of the YTLR software tool helps researchers to obtain the major experimental methods and conditions in the full texts. Since researchers may need to understand the experimental methods and conditions for a given TFB or TFR article, we further fined-tuned a BioBERT-based question–answering network that can help identify the mentioned experiments within the "Materials and Methods" sections of the given full-texts. In the fined-tuned question–answering network, it will try to extract the first occurrence of answers to the question: What are the primary experiments and conditions for the given research? The question–answering network is trained and validated on the YEASTRACT-curated data of paragraphs with known experiment conditions and measurement methods. The performance of standard question–answering models is usually estimated by the strict accuracy (sACC) value. sACC is defined by counting the percentage of predictions that have more than 50% overlap with the true answer sentence segments [30]. As a result, the fine-tuned question–answering network can achieve a test sACC value of 0.35, which indicates that the network can help extract the experiment information within some articles. Hence, the possible experimental conditions and methods are also provided for the users. Last, the YTLR software tool also supports GPUs (graphics processing units) to accelerate the pipeline. When a GPU is set up on the system, YTLR can utilize the GPU to speed up the computation. And a simple one-line command to the YTLR software tool can automatically finish all these features. The README file with step-by-step installation instructions and the YTLR tool can be downloaded from the URL in the ”Data Availability” section.

3.4. Issues related to YTLR

YTLR is a two-staged auto literature reader pipeline for TF-gene transcriptional association extraction. In the training steps of the deep learning networks, the quality of the ground truth positive samples and negative sets usually determines the general performance of the models [31]. In constructing the sample-bootstrapping ensemble models in YTLR Phase I, the negative sets were collected from the deposited yeast articles in PubMed that do not contain any YEASTRACT-curated TFB or TFR evidence. However, there might be articles describing TFB or TFR evidence that were missed by human curators, contributing to incorrectly labeled negative samples. The percent of missed papers that actually describe TFB or TFR evidence can be estimated using false-negative rates (FNR):

$$\mathrm{FNR} = \frac{FN}{FN + TP},$$

where TN and FN are defined in the ”Performance of YTLR” section. In YTLR Phase I, the TFB/TFR article identification networks result in around 9.5%/8.4% FNRs on the test sets. Hence YTLR may miss around 9% more transcriptional regulation-related papers than human curators when trying to save tremendous numbers of human resources.

It is worth noticing that in YTLR Phase II, only TF-gene binding/regulatory associations with evidence discussed in the full texts or abstracts can be extracted. In some yeast transcriptional regulation-related research works, the novel TF-gene pairs can only be inferred from reanalyzing or re-interpreting the experimental results from the supplementary files or data deposited elsewhere. Since these TF-gene pairs lack the proper sentence descriptions within the articles, YTLR ignores these TF-gene pairs. Currently, YTLR aims to provide automated TF-gene transcriptional association extraction only from the literature texts. Recognizing these pure experimental data-inferred TF-gene pairs requires completely different reasoning methods other than natural language understanding. Developing experimental data deduction models can be a future topic in biomedical literature machine understanding.

Currently, YTLR aims to help automatically identify TF-gene binding pairs and regulatory associations for the yeast species since large-scale manually curated TF-gene transcriptional associations are only available in *Saccharomyces cerevisiae*. However, the designed pipeline for generating auto literature readers is general and applicable for different species. Transfer learning techniques [32] can be further applied to these networks to obtain the auto literature readers for other species. Therefore, YTLR can be updated to support the auto-curation of TF-gene transcriptional associations in species other than yeast when manually curated transcription-related literature datasets are available for those species in the future.

4. Conclusions

We constructed a two-phased machine literature reader pipeline called YTLR (Yeast Transcription-regulation Literature Reader) to automate the extraction of yeast TF-gene transcriptional associations from the literature. In YTLR Phase I, articles with TFB evidence and TFR results are identified. Then in YTLR Phase II, TF-gene pairs in a TFB article are recognized as TF-gene binding pairs or random pairings within the article, and TF-gene pairs from a TFR paper are summarized to be TF-gene regulatory associations or random pairings within the paper. YTLR is demonstrated to achieve high AUC values in identifying articles with TFB and TFR evidence and in mining out detailed TF-gene binding pairs and regulatory associations from TFB articles and TFR papers. Moreover, YTLR outperforms previous models and baseline methods in both tasks and shows a modest miss rate compared with human curators. By building a software tool to facilitate large-scale literature TF-gene transcriptional relation summarization, we believe that YTLR can speed up the knowledge accumulation in TF-gene transcriptional regulation research for the community.

Data Availability

The YTLR literature TF-gene association extraction pipeline and the supplementary file can be downloaded at https://github.com/cobisLab/YTLR/.

CRediT authorship contribution statement

Tzu-Hsien Yang: Conceptualization, Investigation, Supervision, Project administration, Formal analysis, Writing - original draft, Writing - review & editing. **Chung-Yu Wang:** Investigation, Software, Visualization, Writing - original draft. **Hsiu-Chun Tsai:** Investigation, Software, Visualization, Writing - original draft, Writing - review & editing. **Ya-Chiao Yang:** Investigation, Software, Visualization, Writing - original draft, Writing - review & editing. **Cheng-Tse Liu:** Software.

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.

Acknowledgments

The authors would like to thank Chi-Fen Liao for helping with the preliminary model evaluation. This work was supported by National Cheng Kung University and the National Science and Technology Council of Taiwan (MOST 107-2218-E-390-009-MY3, MOST 110-2222-E-006-017, and MOST 111-2221-E-006-231).

References

[1] Yang T-H, Wang C-C, Hung P-C, Wu W-S. cisMEP: an integrated repository of genomic epigenetic profiles and cis-regulatory modules in Drosophila. BMC Syst Biol 2014;8(4):1–11.

[2] Hu Z, Killion PJ, Iyer VR. Genetic reconstruction of a functional transcriptional regulatory network. Nature Genet 2007;39(5):683–7.

[3] Hardison RC, Taylor J. Genomic approaches towards finding cis-regulatory modules in animals. Nature Rev Genet 2012;13(7):469–83.

[4] Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO. Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 2000;11(12):4241–57.

[5] Sarda S, Hannenhalli S. High-throughput identification of cis-regulatory rewiring events in yeast. Mol Biol Evolution 2015;32(12):3047–63.

[6] Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne J-B, Reynolds DB, Yoo J, et al. Transcriptional regulatory code of a eukaryotic genome. Nature 2004;431(7004):99–104.

[7] Yang T-H. Transcription factor regulatory modules provide the molecular mechanisms for functional redundancy observed among transcription factors in yeast. BMC Bioinformatics 2019;20(23):1–16.

[8] Yang T-H, Wang C-C, Wang Y-C, Wu W-S. YTRP: a repository for yeast transcriptional regulatory pathways. Database 2014.

[9] Teixeira MC, Monteiro PT, Palma M, Costa C, Godinho CP, Pais P, Cavalheiro M, Antunes M, Lemos A, Pedreira T, et al. YEASTRACT: an upgraded database for the analysis of transcription regulatory networks in saccharomyces cerevisiae. Nucl Acids Res 2018;46(D1):D348–53.

[10] Yang P, Oldfield A, Kim T, Yang A, Yang JYH, Ho JW. Integrative analysis identifies co-dependent gene expression regulation of BRG1 and CHD7 at

distal regulatory sites in embryonic stem cells. Bioinformatics 2017;33 (13):1916–20.

[11] Stuckey S, Storici F. Gene knockouts, in vivo site-directed mutagenesis and other modifications using the delitto perfetto system in Saccharomyces cerevisiae. Methods Enzymol 2013;533:103–31.

[12] Yang T-H, Wu W-S. Inferring functional transcription factor-gene binding pairs by integrating transcription factor binding data with transcription factor knockout data. BMC Syst Biol 2013;7(6):1–14.

[13] Sopko R, Huang D, Preston N, Chua G, Papp B, Kafadar K, Snyder M, Oliver SG, Cyert M, Hughes TR, et al. Mapping pathways and phenotypes by systematic gene overexpression. Mol Cell 2006;21(3):319–30.

[14] Gitter A, Siegfried Z, Klutstein M, Fornes O, Oliva B, Simon I, Bar-Joseph Z. Backup in gene regulatory networks explains differences between binding and knockout results. Mol Syst Biol 2009;5(1):276.

[15] Björne J, Ginter F, Pyysalo S, Tsujii J, Salakoski T. Complex event extraction at PubMed scale. Bioinformatics 2010;26(12):i382–90.

[16] Malard F, Wulff-Fuentes E, Berendt RR, Didier G, Olivier-Van Stichelen S. Automatization and self-maintenance of the O-GlcNAcome catalog: a smart scientific database. Database 2021:baab039.

[17] Bugnon LA, Yones C, Raad J, Gerard M, Rubiolo M, Merino G, Pividori M, Di Persia L, Milone DH, Stegmayer G. DL4papers: a deep learning approach for the automatic interpretation of scientific articles. Bioinformatics 2020;36 (11):3499–506.

[18] Holtzapple E, Telmer CA, Miskov-Zivanov N. FLUTE: Fast and reliable knowledge retrieval from biomedical literature. Database 2020.

[19] Sänger M, Leser U. Large-scale entity representation learning for biomedical relationship extraction. Bioinformatics 2021;37(2):236–42.

[20] Rodríguez-Penagos C, Salgado H, Martínez-Flores I, Collado-Vides J. Automatic reconstruction of a bacterial regulatory network using natural language processing. BMC Bioinform 2007;8(1):1–11.

[21] Novichkova S, Egorov S, Daraselia N. MedScan, a natural language processing engine for MEDLINE abstracts. Bioinformatics 2003;19(13):1699–706.

[22] Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, et al. SGD: Saccharomyces genome database. Nucl Acids Res 1998;26(1):73–9.

[23] Laza R, Pavón R, Reboiro-Jato M, Fdez-Riverola F. Evaluating the effect of unbalanced data in biomedical document classification. J Integrative Bioinform 2011;8(3):105–17.

[24] Gessert N, Nielsen M, Shaikh M, Werner R, Schlaefer A. Skin lesion classification using ensembles of multi-resolution efficientnets with meta data. MethodsX 2020;7:100864.

[25] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234–40.

[26] Yang T-H, Wang C-Y, Tsai H-C, Liu C-T. Human IRES Atlas: an integrative platform for studying IRES-driven translational regulation in humans. Database 2021.

[27] Yang T-H, Yang Y-C, Tu K-C. regCNN: identifying Drosophila genome-wide cis-regulatory modules via integrating the local patterns in epigenetic marks and transcription factor binding motifs, Computational and Structural. Biotechnol J 2022;20:296–308.

[28] Burns GA, Li X, Peng N. Building deep learning models for evidence classification from the open access biomedical literature. Database 2019.

[29] Valenzuela-Escárcega MA, Babur Ö, Hahn-Powell G, Bell D, Hicks T, Noriega-Atala E, Wang X, Surdeanu M, Demir E, Morrison CT. Large-scale automated machine reading discovers new cancer-driving mechanisms. Database 2018.

[30] Xu G, Rong W, Wang Y, Ouyang Y, Xiong Z. External features enriched model for biomedical question answering. BMC Bioinformatics 2021;22(1):1–19.

[31] Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nature Genetics 2019;51(1):12–8.

[32] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Know Data Eng 2009;22(10):1345–59.