DESSO-DB: A web database for sequence and shape motif analyses and identification

Xiaoying Wang a,1, Cankun Wang b,1, Lang Li b, Qin Ma b, Anjun Ma b,a, Bingqiang Liu a,*

a Department of Mathematics, Shandong University, Shandong, China
b Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH 43210, USA

ABSTRACT

Cis-regulatory motif (motif for short) identification and analyses are essential steps in detecting gene regulatory mechanisms. Deep learning (DL) models have shown substantial advances in motif prediction. In parallel, intuitive and integrative web databases are needed to make effective use of DL models and ensure easy access to the identified motifs. Here, we present DESSO-DB, a web database developed to allow efficient access to the identified motifs and diverse motif analyses. DESSO-DB provides motif prediction results and visualizations of 690 ENCODE human Chromatin Immunoprecipitation sequencing (ChIP-seq) data (including 161 transcription factors (TFs) in 91 cell lines) and 1,677 human ChIP-seq data (including 547 TFs in 359 cell lines) from Cistrome DB using DESSO, which is an in-house developed DL tool for motif prediction. It also provides online motif finding and scanning functions for new ChIP-seq/ATAC-seq datasets and downloadable motif results of the above 690 DECODE datasets, 126 cancer ChIP-seq, 55 RNA Crosslinking-Immunoprecipitation and high-throughput sequencing (CLIP-seq) data. DESSO-DB is deployed on the Google Cloud Platform, providing stabilized and efficient resources freely to the public. DESSO-DB is free and available at http://cloud.osubmi.com/DESSO/

1. Introduction

De-novo motif prediction from given genomic sequences (e.g., promoters and enhancers) helps the identification of transcription factor binding sites (TFBSs) and the elucidation of gene regulatory networks [1]. Deep learning (DL) models have been considered potent methods for motif finding and TFBS identification from Chromatin Immunoprecipitation sequencing (ChIP-seq) data (Alipanahi et al., 2015) and Crosslinking-Immunoprecipitation and high-throughput sequencing (CLIP-seq) data, achieving better performances than traditional methods, such as gkmSVM and MEME-ChIP Zhang [20]. However, considering the required computational resources and programming skills, utilizing motif-finding DL tools is not easy. As an alternative option, single cell data can better reflect the heterogeneity between cells and complex disease regulatory mechanisms [16] and Cleavage Under Target & Release Using Nuclease (CUT&RUN) data [4] outperforms ChIP-seq protocols in resolution, signal-to-noise ratio, and required sequence depth and does not require cross-linking [1]. However, there is no motif identification tool specifically designed for single-cell data at present.

We have previously developed an in-house tool named DESSO Yang [18], which was benchmarked as the best tool with the highest overall score for motif finding and TFBS identification among 20 DL tools Zhang [20]. DESSO improves the identification and structural analysis of TF binding sites, by integrating the complexities of DNA binding into a DL framework. To enable easy access to DESSO functions and results, we developed DESSO-DB on the Google Cloud Platform, collecting motif-finding results of DESSO from 690 ENCODE ChIP-seq datasets (covering 161 TFs in 91 cell lines) and 1,677 Cistrome DB ChIP-seq datasets (covering 547 TFs in 359 cell lines). We used the online motif scan and de novo motif finding functions of DESSO for new DNA sequences and ChIP-seq data. To test whether DL tools developed for bulk ChIP-seq data can also be used for single-cell CUT&RUN data, we applied DESSO to 172 single-cell CUT&RUN samples. We found 240 representative motifs in the 172 cells, including 65, 17, and 13 motifs that were uniquely identified in CTCF, SOX2, and NANOG data, respectively. We also found shared motifs and revealed potential TF co-regulatory mechanisms. Moreover, DESSO-DB provided downloadable results for the identified motifs on 55 CLIP-seq datasets Yang [18,19], and 126 cancer-related ChIP-seq datasets. All of these results are downloadable on DESSO-DB.
2. Materials and methods

2.1. Data collection

For ChIP-seq data, the 690 ChIP-seq datasets were analyzed in the original DESSO paper from the ENCODE Analysis Database at UCSC (https://genome.ucsc.edu/ENCODE/downloads.html), which included 161 TFs and 91 human cell types. The peaks were obtained from the SPP peak caller and de-noised by the Irreproducible Discovery Rate Li [8] based on signal reproducibility among biological replicates. The peaks were sorted in decreasing order of corresponding signal scores in each data file. For CLIP-seq data, 55 CLIP-seq datasets were collected from references [12] and a fixed length of 101 bps was used as input [11]. The Cistrome DB, database includes 11,729 human datasets, which pass all quality control (sequence quality, mapping quality, library complexity, ChIP enrichment, signal noise ratio, and regulatory region). ATAC-seq datasets and histone modification datasets were removed. To utilize limited resources, increase the effectiveness of the database, and consider the conservative motif, we reserved all cell lines and TFs without repetition rather than introducing all datasets without distinction. For each cell line and TF, we randomly selected a representative dataset. Therefore, the 1,677 ChIP-seq datasets were downloaded, covering 547 TFs and 359 cell types. All datasets can be found in Supplementary Data S1-3.

2.2. Data processing

For "bed" format data, bedtools (2.29.2) was used to covert location information (".bed" format) to sequence information (".fasta" format). For the ChIP-seq datasets and CLIP-seq datasets, we pruned the original peaks to a fixed length (101 bps) using formula (1). The position of the processed peaks in the chromosomes can be identified by:

\[
\text{position} = \left[ \frac{\text{origin} + \text{end}}{2} \right] - 50, \quad \left[ \frac{\text{origin} + \text{end}}{2} \right] + 50 \# 
\]

where origin is the start position of the original peak, end is the end position of the original peak position. After pruning, redundant peaks can be generated but were removed in this experiment, and bedtools v2.29.2 was employed to acquire the pruned sequences [14]. Binary vectors were required as input for DESSO. Each input sequence was converted to an encoded matrix \( M = L \times 4 \), i.e. \( A = [0, 0, 0, 1], \; G = [0, 1, 0, 0], \; C = [0, 0, 1, 0], \; T = [0, 0, 0, 1] \), where \( L \) is the length of the input sequence Jinyu [6]. For 172 single-cell CUT&RUN datasets, the length of peaks in each dataset was different, so we pruned the original peaks to a fixed length (1001 bps) using formula (1). The pruned sequences were encoded to one-hot cods in the same way as ChIP-seq and CLIP-seq in the previous paragraph. All datasets only contained positive samples (peak sequences). Negative samples were generated by selecting the same number of 101 bps sequences from GENCODE with the same Guanine-Cytosine (GC) content Harrow [5]. Positive and negative sequences had the same GC content without any overlapping peaks. In our experiments, each positive sample was labeled as ‘1’, and each negative sample was labeled as 0.

2.3. Motif prediction using DESSO

DESSO split the input preprocessed data into three parts: training data, validation data, and test data. The top 500 odd peaks in each dataset were selected to test the models, while 80% of the rest of the peaks were selected to train the model, and the remaining peaks were used for validation. DESSO is composed of a convolutional neural network (CNN) for classification to extract motif patterns from preprocessed data and a statistical model for optimizing motif instances selection.

The CNN is composed of a convolution layer, a max pooling layer, a fully connected layer, and an output layer. The convolution layer is defined as in formula (2):

\[
A_k = \text{Relu} (\text{Conv}_k (M))
\]

where \( k = 1, \ldots, K, \; K \) is the number of convolutional filters and is set to 16. \( M \) is the input data obtained by data preprocessing. Conv is the convolution layer, and Relu is an activation function, which is defined as \( \text{Relu}(x) = \max(0, x) \). Then, the max pooling layer was used to reduce the dimension of \( M \) and maintain the motif invariance. The max pooling layer is defined as in formula (3):

\[
P = \text{MaxPool}(A_k)
\]

and the \( P \) is fed into a fully connected layer, which is defined as in formula (4):

\[
Z = \text{Relu} (\langle W_1, P + b_1 \rangle)
\]

where \( W_1 \) is the weight in the fully connected layer and \( b_1 \) is the bias in the fully connected layer. \( Z \) is the output of the fully connected layer. Finally, we fed \( Z \) into the output layer using formula (5):

\[
y = \text{sigmoid} (W_2 Z + b_2)
\]

where \( W_2 \) is the weight in the output layer and \( b_2 \) is the bias in the output layer. sigmoid is an activation function to limit the range of predicted results rant 0–1. sigmoid\( (x) = \frac{1}{1 + e^{-x}} \). \( y \) is the predicted result.

For each dataset, we trained the CNN model by minimizing the negative loglikelihood loss using formula (6):

\[
\text{Loss} = \frac{1}{N} \text{NLL}(y, l) + \lambda \| \alpha \|_2^2
\]

where NLL is a negative log likelihood function, \( \alpha \) is a regularization parameter, and \( \| \cdot \|_2 \) indicates the \( L_2 \) norm. \( y \) is the predicted result and \( l \) is the true label.

After the training process, the binomial distribution was used to optimize motif instance \( \Omega(S) \), in which \( S \) is the binding site signal matrix from the CNN model. We denoted a random variable as \( X \) which represents the number of the predicted signal site containing the background sites, so which \( f(x) = P(X = x) \) is the probability function. DESSO assumed that \( f(x) \) can be regarded as a binomial distribution \( X \overset{\text{Binomial}}{\sim}(n, p) \), where \( n \) is the background sequence number and \( p = \frac{r}{n} \). The optimized motif instances were obtained with a p-value < 1 \times 10^{-4}.
A set of aligned activated sequences was obtained based on the optimized motif instances. A position weight matrix (PWM) [15] can be generated in rows as the four nucleic acid types (A, T/U, C, and G), in columns as sequence positions, and an element as nucleotide occurrences in the corresponding position [2]. The PWM was computed based on the set of aligned activated sequences. The PWM was aligned as the motif profiles and visualized using WebLogo [2]. We identified the underlying TFs for the motifs by querying the HOCOMOCO v11 database through the TOMTOM v5.1.0 tool. The predicted TF binding site position files were first merged into a bigBed track file. We retrieved the genome browser track file from JASPAR, which stores all known TF binding sites of each TF. The p-value score was calculated as $-\log_{10}(p) \times 100$ in JASPAR, where 0 corresponds to a p-value of 1 and 1000 corresponds to a p-value < $10^{-10}$. We matched each binding site in the bigBed track file with the most significant TF.

2.4. Single cell CUT&RUN data analysis

To test whether DESSO is equally effective for single-cell data, we applied DESSO to 172 mouse embryonic stem cells of single cell CUT&RUN sequencing, including 120 cells for CCCTC-Binding factor (CTCF), 26 cells for Sex Determining Region Y Box 2 (SOX2), and 26 cells for Nanog homeobox (NANOG) to identify TF co-regulatory patterns of specific genes [4]. The cells contained TFBSs with lengths of 1–120 bps for each single-cell CUT&RUN dataset. CUT&RUN datasets can be found in Supplementary Data S4. For finding known motifs from motif instances, TOMTOM was used to carry out the pair-wise motif comparison with the known motifs in the HOCOMOCO database. The motif instances were defined as motifs if the comparison p-value < 1e-2. To better describe the motifs, we selected matched motifs with a minimum p-value from HOCOMOCO Kulakovskiy [7] as representative motifs. We performed motif merging with TOMTOM. TOMTOM defined the similarity score between query and target motifs and converted the scores into p-values Gupta [3]. The Pearson correlation coefficient was used for motif similarity calculation in TOMTOM, which was first introduced by Pietrokovski [13]. To explore the relations between the motif co-enrichment and TF co-regulation, we mapped the five motifs shared by SOX2 and NANOG data to the upstream 5 kb region of Pou5f1 using the mouse reference genome (GRCm38/mm10) and compared the mapping results with the peak region in the same region. The potential TF target genes were determined through the matching of identified motifs to the upstream promoter region via Cistrome-GO Zheng [21].

2.5. Database construction

In consideration of the power of cloud computing, we host DESSO-DB on the Google Cloud Platform, which is freely available to public users. The backend server was organized and deployed using Nextflow to ensure flexibility and scales efficient scaling on the cloud. Once a query job is submitted, it will be placed in the Cloud SQL database, dispatched to cloud nodes at scale, and analyzed by the DESSO workflow Yang [18]. Finally, the results will be returned and available for visualization and download. A unique job ID will be generated and emailed to the user when the analysis is complete if an email address is provided.

Fig. 1. Overview of the webserver. (A) The data table for 2,367 ChIP-Seq datasets with 410 cell lines and 576 TFs. Each entry indicates the number of datasets derived from a specific cell line and a specific TF. The MAFF Transcription Factor F (MAFF) dataset in the HepG2 cell line was highlighted as an example. (B) The identified MAFF sequence and shape motif using DESSO, with links comparing their position weight matrix (PWM) to the existing motif databases. (C) Snapshot of the MAFF sequence profile (i.e., motif instances information). (D) Line chart for the per-nucleotide vertebrate motif conservation and the ± 50 bps flanking regions within the HepG2 cell line. (E) The line chart for the corresponding mean motif value and the ± 50 bps flanking regions within the HepG2 cell line. (F) The occurrence of the MAFF motif in the corresponding ChIP-seq peaks, which are ranked by peak signal, and an enrichment score of the motif in its corresponding CHIP-seq peaks. (G) The UCSC Genome Browser track hub displays genome-wide predicted binding sites for each binding profile in DESSO-DB.
3. Results

3.1. Overview of DESSO-DB

DESSO-DB provides the following three major functions to explore identified motifs and predict new motifs. (i) We provided a grid table for users to check motif results of the 2,367 ChIP-seq data identified by DESSO (Fig. 1A). Detailed interpretations are listed (Supplementary Example S1), including TF information, motif sequence profiles and logo, and shape motif-related flanking region. All motif results can also be accessed by searching for specific TF names, UCSC accession numbers, cell lines, or job IDs. (ii) The motif scanning function searches for all motif instances of a query motif selected from DESSO results in given genomic sequences (Supplementary Examples S2 and S3). (iii) De novo motif finding identifies a set of statistically significant motifs (if any) in a set of provided promoters using the trained DESSO model (Supplementary Example S4).

Fig. 2. Single-cell CUT&RUN analysis. (A) An upset plot of the shared and unique motifs among CTCF, NANOG, and SOX2 datasets. (B) Motifs are found in each cell, and motifs identified are denoted as blue. (C) Co-enrichment of SOX2 and NANOG motifs in the upstream regulatory regions of genes Pou5f1 and Pgk1. (D) The regulatory network of SOX2 and NANOG is inferred from motif co-enrichment. The pink nodes represent the genes shared by NANOG and SOX2 with higher potentials to be co-regulated by SOX2 and NANOG.
Four major visualizations were provided online to display the above functions. (i) Each of the identified motifs was annotated by comparisons (using TOMTOM) with existing databases, such as JASPAR, Mathelier [10], and HOCOMOCO Kulakovskiy [7] (Fig. 1B). Specifically, a shape motif profile had ± 50 bps flanking regions added using a bold orange curve (i.e. Minor Groove Width, Propeller Twist, Helix Twist, and Roll) features computationally derived from DNA sequences by Monte Carlo simulation [22]. The two boundary curves of the blue region represented the upper and lower boundaries of the shape features in the corresponding motif instances. (ii) The identified motif instances and locations were presented in a well-organized table (Fig. 1C). (iii) A line chart for the per-nucleotide vertebrate conservation and the corresponding mean value of motifs was available (Fig. 1D and E). Each figure had ± 50 bps flanking regions added within the cell line. (iv) The enrichment score for the identified motifs was provided (Fig. 1F). The red curve indicated the enrichment score on the corresponding ChIP-seq peaks. The vertical black lines indicated the presence of ChIP-seq peaks that contained at least one TFBS. (v) To boost the usability of the DESSO-DB, we provided DESSO-DB predictions as genome tracks in the Download link and displayed genome-wide predicted binding sites for each binding profile in the UCSC Genome Browser track hub (Fig. 1G). All motif results of the 20 DL tools can be downloaded on DESSO-DB. We also provide downloadable motif results of DESSO on the 690 ENCODE ChIP-seq datasets, 55 CLIP-seq datasets, 126 cancer-related ChIP-seq datasets, and 1,677 Cistrome DB datasets (Supplementary Example S3).

3.2. Motif analysis on single-cell CUT&RUN data reveals potential TF co-regulatory mechanisms

We identified 118, 124, and 205 motifs from SOX2, NANOG, and CTCF data, respectively. Then, 477 motifs were merged by pairwise motif similarity comparison. Eventually, 240 representative motifs remained in the 172 cells for the following analysis, including 65, 17, and 13 motifs that were uniquely identified in CTCF, SOX2, and NANOG data, respectively (Supplementary Data S5). Additionally, 62 motifs were shared by SOX2, NANOG, and CTCF (Group 1), 44 motifs were shared by NANOG and CTCF (Group 2), 34 motifs were shared by SOX2 and CTCF (Group 3), and five motifs were shared by SOX2 and NANOG (Group 4) (Fig. 2A and B).

With the observed motifs being shared by different CUT&RUN data, we explored the relations between the motif co-enrichment and TF co-regulation. It has been reported that SOX2 and NANOG co-regulate the expression of the Pou5f1 gene and maintain pluripotency in human embryonic stem cells Wang [17]. We observed that 88% of SOX2 cells and 88% of NANOG cells had motif enrichment in the same corresponding peak region, indicating that the motif co-enrichment of SOX2 and NANOG lead to the co-regulation of Pou5f1 (Fig. 2C). Motif co-enrichment of SOX2 and NANOG was also found in the 3 kb upstream regulatory region of Pgk1, which suggests that Pgk1 can also be co-regulated by SOX2 and NANOG.

To further investigate how SOX2 and NANOG co-regulate their downstream target genes, we constructed a gene regulatory network for each of the TFs (Fig. 2D). We identified two groups of 37 genes that were solely regulated by SOX2 and NANOG as well as 13 genes that are regulated by both TFs, including Pgk1 and Pou5f1. Interestingly, both Pou5f1 and Pgk1 were reported as marker gene candidates for the monitoring of embryo development stages Mamo [9]. Therefore, we reasoned that the rest of the 13 genes co-regulated by SOX2 and NANOG were also potential marker gene candidates for the mouse embryonic stem cells. Our results suggest that the existing DL methods are suitable for analyzing single-cell CUT&RUN data and can reveal potential TF co-regulatory patterns of potential marker gene candidates at the single-cell level.

4. Conclusions

Motif identification and analyses provide a solid foundation to infer gene regulatory networks encoded in a genome. Our in-house DESSO and benchmarking analysis showed considerably better performances than traditional methods and other DL tools Yang [18,20]. We developed DESSO-DB to provide easy access to the results and functions of DESSO and other DL tools. We believe that DESSO-DB is a highly useful and user-friendly platform for motif identification and analyses, and that it benefits the genomic research community.

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.

Funding

This work was supported by National Key R&D Program of China (2020YFA0712400), National Nature Science Foundation of China (NSFC, 11931008), and Shandong University multidisciplinary research and innovation team of young scholars (2020QNQT017).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.06.031.

References

[1] Bushweller JH. Targeting transcription factors in cancer — from undruggable to reality. Nat Rev Cancer 2019;19(11):611–24.
[2] Crooks GE et al. Weblogo: a sequence logo generator. Genome Res 2004;14 (6):1188–90.
[3] Gupta S et al. Quantifying similarity between motifs. Genome Biol 2007;8:2.
[4] Hainer, S.J., et al. Profiling of pluripotency factors in single cells and early embryos. Cell 2019;175(5):1319-1329. e1311.
[5] Harrow J et al. GENCODE: The reference human genome annotation for The ENCODE Project. Genome Res 2012;22(9):1760–74.
[6] Jinju Y et al. Prediction of regulatory motifs from human Chip-sequencing data using a deep learning framework. Nucleic acids research 2019:15:15.
[7] Kulakovskiy IV et al. HOCOMOCO: towards a complete collection of transcription factor binding profiles. Nucleic Acids Res 2016;44(D1):D110–115.
[8] Li QH et al. Measuring Reproducibility of High-Throughput Experiments. Ann Appl Stat 2011;5(3):1752–79.
[9] Mamo S et al. Expression profiles of the pluripotency marker gene Pou5f1 and validation of reference genes in rabbit oocytes and preimplantation stage embryos. BMC Mol Biol 2008;9(1):67.
[10] Mathelier A et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res 2016;44(D1):D110–115.
[11] Pan X, Shen H-B. RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. BMC Bioinf 2017;18(1):136.
[12] Pan X, Shen H-B. Learning distributed representations of RNA sequences and its application for predicting RNA-protein binding sites with a convolutional neural network. Neurocomputing 2018;305:51–8.
[13] Pietrokovski S. Searching databases of conserved sequence regions by aligning protein multiple-alignments. Nucleic Acids Res 1996;24(19):3836–45.
[14] Quinlan, A.R. and Hall, I.M. BEDTools. Current Protocols in Bioinformatics 2016;47(12.11.11):12.11.11.
[15] Rahul S, Raya K. Dinucleotide Weight Matrices for Predicting Transcription Factor Binding Sites: Generalizing the Position Weight Matrix. PLoS ONE 2010;5(3):e9722.
[16] Stuart T, Satija R. Integrative single-cell analysis. Nat Rev Genet 2019;20 (5):257–72.
17. Wang Z et al. Distinct lineage specification roles for NANOG, OCT4, and SOX2 in human embryonic stem cells. Cell Stem Cell 2012;10(4):440–54.

18. Yang J et al. Prediction of regulatory motifs from human Chip-sequencing data using a deep learning framework. Nucleic Acids Res 2019;47(15):7809–24.

19. Yi HC et al. Learning distributed representations of RNA and protein sequences and its application for predicting IncRNA-protein interactions. Comput Struct Biotechnol J 2020;18:20–6.

20. Zhang S et al. Assessing deep learning methods in cis-regulatory motif finding based on genomic sequencing data. Briefings Bioinf 2021.

21. Zheng R et al. Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. Nucleic Acids Res 2018;47:D729–D735.

22. Zhou, T., et al. DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic Acids Res 2013;41(Web Server issue):W56–62.