Prediction of Transcription Factor Binding Sites using TFBIND tool

BMS 320, Spring 20

Muhammad Salah Elsadany
Computational Biology and Genomics, Biomedical Sciences Program
University of Science and Technology, Zewail City
Giza, Egypt
s-muhammadelsadany@zewailcity.edu.eg

1. Introduction

Transcription is considered to be the first step in gene expression inside the cell. It involves copying the DNA sequence of a certain gene to make an RNA. Then, this RNA strand will be translated to a functional product, called protein. The whole process can be performed using certain enzymes called RNA Polymerases [1]. These enzymes link nucleotides to form an RNA strand by using the original DNA strand as a template for that. This RNA strand, called transcript, carries all needed information to build a polypeptide chain. The transcription process happens in 3 main stages: initiation, elongation, and termination [1]. The initiation part starts by the binding of the RNA polymerase to the promoter region of the gene of interest needed to be transcribed, found near the beginning of that gene. That promoter region differs for each gene and not the same to increase the specificity needed for each gene transcription. Once the RNA polymerase is bound, it starts separating the DNA strands, providing the single-stranded template needed for the transcription process to proceed to the elongation part [2].

In bacteria, the RNA polymerase attaches right to the promoter of needed genes and it can be regulated by transcription factors. In eukaryotes, there is an extra step that the RNA polymerase can attach to the promoter only with the help of proteins called general transcription factors. Transcription factors are proteins that are involved in the process of transcribing a gene. They include a wide number of proteins that initiate and regulate the transcription process of genes. However, the RNA polymerase is not included among that group named transcription factors [3]. The transcription factors can be categorized into two main groups: activators or enhancers and terminators or repressors. The activators may help the general transcription factors and the RNA polymerase bind to the promoter. On the other hand, repressors repress the transcription process as they may get in the way of the general transcription factors or RNA polymerase, making it difficult for them to bind to the promoter or begin transcription. These transcription factors have DNA-binding domains that give them the ability to bind tightly to specific sequences of the gene called enhancer or promoter sequences [3]. These regulatory sequences can be thousands of base pairs upstream or downstream from the gene being transcribed. However, they can also be found in other parts of the DNA, sometimes very far away from the promoter region. That action of any transcription factors allows for unique expression

for each gene in different cell types and different environments.

The target model in this project is Saccharomyces cerevisiae that lives in boom and bust nutritional environments. That organism has evolved to rapidly cope with any environment changes while preserving intracellular homeostasis. The targeted system in this project is the nitrogen regulation part and coping to different nitrogen levels in different environments. Target of Rapamycin Complex 1 (TorC1), is a nitrogen-responsive regulator [4]. TorC1 is activated by excess nitrogen in the environment and downregulated by limiting nitrogen. Two of TorC1's many downstream targets are Gln3 and Gat1 whose localization and function are Nitrogen Catabolite Repression-(NCR-) sensitive [4]. They are both GATA-family transcription activators. The nitrogen-regulated genes are expressed upon nitrogen limitation. The TorC1 inhibits expression of nitrogen-regulated genes by sequestering the GATA-binding transcription factors GLN3 and GAT1 in the cytoplasm [5]. In the presence of a good nitrogen source, GLN3 is TorC1-dependently phosphorylated and therefore bound to the cytoplasmic URE2 protein. Upon nitrogen limitation, GLN3 is dephosphorylated by type 2A-related phosphatase SIT4, released from URE2 and transferred to the nucleus where target genes are activated [5]. Therefore, GAT1 is required for the activation of transcription of a number of genes in response to the replacement of glutamine by glutamate as source of nitrogen. On the other hand, the second targeted gene in this project, Dal80, is a negative regulator of genes in multiple nitrogen degradation pathways [6]. The aim of this project is to get the transcription factors binding sites or domains inside each of these two genes, Dal 80 and GAT1, using the TFBIND online tool.

2. Problem Definition

Accurate prediction of transcription factor-DNA interaction is an effective indicator for estimating interactions between TFs, predicting DNA gene promoter, estimating gene function by analyzing the organizational specificity of its upstream regulatory region, and rebuilding genetic networks. Activation or suppression of transcription is indeed one of the essential levels of gene regulation. Documentation on experimentally validated workable TFBSs is limited and therefore there is a need for accurate prediction of TFBSs for gene annotation and applications. There is therefore a desire for computational techniques to predict cell type-specific TF binding with good accuracy.

3. Related Work and Survey

There are plenty of tools available for doing the same project target and of them has its own advantages and techniques. These tools include DOOR2, TFmiR, CONREAL, and PlnTFDB … etc. These are all free available online tools. The DOOR2 is Database of prOkaryotic OpeRons and it offers a high-performance web service for online operon prediction, and an intuitive genome browser to support visualization of user-selected data. TFmiR is mainly for deep and integrative analysis of combinatorial regulatory interactions between transcription factors, microRNAs and target genes that are involved in disease pathogenesis. CONREAL allows identification of transcription factor binding sites that are conserved between two sequences. On the other hand, there are different available R packages that do the same function with slight differences and more analysis added to the output. These packages may include generegulation and TFBSTools. In this project, there will be a comparison between the used online tools,

TFBIND, with one of the R packages which is generegulation. The generegulation package works mainly on finding candidate TF binding sites in DNA sequence using the model organism Saccharomyces cerevisiae which is the same used one in this project [7].

4. Proposed Method

The used tool TFBIND is a free available online tool that helps in getting all transcription factor binding sites from the entered sequence of interest. It's available in [8]. This tool has an advantage over other available online tools because it doesn't have a limit for the number of base pairs of the entered sequence. The building of this tool used 205 vertebrate transcription factors from the database TRANSFAC Ver 3.4. And that gives it a privilege of being more concise and accurate.

The algorithm of this tool works on calculating the matching score between an input sequence and a set of known transcription factor binding sites. In order to achieve this, the tool uses Position Weight Matrices (PWMs) [9] and Bucher's calculating method [10]. The data source for this tool comprised 433 non-redundant vertebrate promoters including viral promoters, from Eukaryotic Promoter Database (EPD) R.50.

For creating a position weight matrix, a Position Frequency Matrix (PFM) is formed first. The position frequency matrix can be constructed by recording the position-dependent frequency of each nucleotide in the DNA sequence that interacted with the transcription factor. For example, the figure below shows a constructed position frequency matrix from [11]. This PFM was done by collecting experimentally validated binding sites from 8 published studies for MEF2 [11].

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

For instance, the first column in the alignment consists of no As, three Cs, two Gs and three Ts, therefore resulting in the corresponding first matrix column {0, 3, 2, 3}. This matrix should be converted into another one which is the row frequency table. This row frequency table could be calculated by dividing each cell of the previous matrix over the total number of letters in the sequence. For example, the first cell of second row, corresponding to Cs at first position, will give = 3/ (14*8) = 0.027 {a}. To convert a position frequency matrix to the corresponding position weight matrix, the frequencies are converted to normalized frequency values on a log-scale. To perform this conversion we can use this formula:

$$W_{b,i} = log_2 \frac{p(b,i)}{p(b)}$$

Where $W_{\{b,\ i\}}$ = position weight matrix value of base b in position i, p(b) = background probability of base b, it can be assumed as 0.25 (4 nucleotides distributed uniformly in background). p(b, i) is

$$p(b,\ i) = \frac{f_{b,i} + s(b)}{N + \sum s(b')}$$

where b' could be {A, C, G, T} ; $f_{\{b,i\}}$ = counts of base b in position i; N = number of sites; p(b,i) = corrected probability of base b in position i and s(b) = pseudocount function. By calculating this for the same given example at {a}, it will give $\log_2$ (0.027) = -1.89. The same procedure will be conducted for the whole matrix to get the position weight matrix at the end. After that, the summation of log odds scores will be

calculated to get a score for the matching sequence.

Using the position weight matrix data, each transcription factor matching score in each position was calculated according to Bucher's method [11]. The Bucher's method is for calculating the binding score at each position on DNA. The previous 2 equations can be simply transferred into this one:

$$W_{b,i} = \ln (P_i (b) + a)$$

Where b refers to the base (b) belonging to {A, T, G, C} at position I of the input sequence, p_i (b) is the probability of each base b at each position I, and (a) is a smoothening parameter (=0.01).

For the input sequence, the binding score is calculated using this formula:

$$x = \sum_{i=1}^{L} W_{b,i}$$

Where L is the consensus length of the motif. To normalize the score, the hypothetical maximum score and minimum score are calculated by these two equations:

$$x_{max} = \sum_{i=1}^{L} \max_b W_{b,i},$$

$$,x_{min} = \sum_{i=1}^{L} \min_b W_{b,i}.$$

Thus, the score of the input sequence is normalized: and the match =

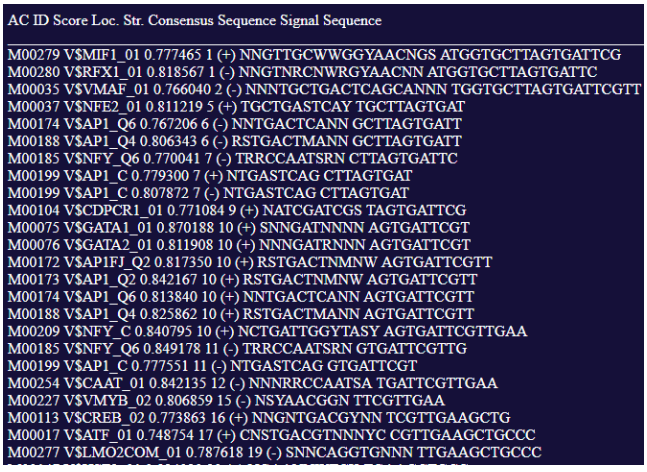
$$\frac{x - x_{\min}}{x_{\max} - x_{\min}}.$$

5. Evaluation

As mentioned earlier, the data set used for evaluating the tool here is 2 main genes to detect their transcription factor binding sites. These two genes are Dal80 and GAT1. Both of the genes' FASTA format were retrieved from NCBI, Gene Database [12, 13]. Both FASTA formats were used as an input for the TFBIND online software tool.

There are two methods to input the data in TFBIND tool: first one is by entering the sequence itself in the text box in a FASTA format, and the second one is by choosing the file and uploading it.

For the Dal80 gene, a sample of the output is shown below:



For the GAT1 gene, a sample of the output is shown below:

AC	ID	Score	Loc.	Str.	Consensus Sequence	Signal Sequence
M00055	V$NMYC_01	0.764294	1	(+)	NNNCACGTGNNN	ATGCACGTTTTC
M00055	V$NMYC_01	0.762165	1	(-)	NNNCACGTGNNN	ATGCACGTTTTC
M00123	V$MYCMAX_02	0.829262	1	(+)	NANCACGTGNNW	ATGCACGTTTTC
M00183	V$MYB_Q6	0.867635	3	(-)	NNNAACKGNC	GCACGTTTTC
M00217	V$USF_C	0.816169	3	(-)	NCACGTGN	GCACGTTT
M00057	V$COMP1_01	0.815210	4	(-)	NNTNWKGATTGRCNRSRANMRRNN	CACGTTTTCTTTCCTTTGCTTTTC
M00127	V$GATA1_03	0.785154	4	(-)	RNSNNGATAANNGN	CACGTTTTCTTTCC
M00160	V$SRY_02	0.783617	6	(-)	NWWAACAAWANN	CGTTTTCTTTCC
M00278	V$LMO2COM_02	0.797844	6	(-)	NMGATANSG	CGTTTTCTT
M00080	V$EVI1_03	0.716461	8	(-)	AGATAAGATAA	TTTTCTTTCCT
M00148	V$SRY_01	0.960465	9	(-)	AAACWAM	TTTCTTT
M00025	V$ELK1_02	0.804102	10	(-)	NNNNCCGGAARYNN	TTCTTTCCTTTGCT
M00074	V$CETS1P54_02	0.857330	10	(-)	NNAMMGGAWRWNN	TTCTTTCCTTTGC
M00003	V$VMYB_01	0.798643	14	(-)	AAYAACGGNN	TTCCTTTGCT
M00195	V$OCT1_Q6	0.804585	14	(-)	NNNNATGCAAATNAN	TTCCTTTGCTTTTCC
M00210	V$OCT_C	0.808562	14	(+)	CTNATTTGCATAY	TTCCTTTGCTTTT
M00062	V$IRF1_01	0.801115	15	(-)	SAAAAGYGAAACC	TCCTTTGCTTTTC
M00063	V$IRF2_01	0.767161	15	(-)	GAAAAGYGAAASY	TCCTTTGCTTTTC
M00160	V$SRY_02	0.799178	15	(-)	NWWAACAAWANN	TCCTTTGCTTTT
M00227	V$VMYB_02	0.801143	15	(-)	NSYAACGGN	TCCTTTGCT
M00072	V$CP2_01	0.815609	18	(-)	GCNMNAMCMAG	TTTGCTTTTCC

Full results were reported in different files attached. The results are shown in a specified order with defined columns. Each of these columns represents different info needed for all hits:

Column1: Transcription factor matrix ID (from TRANSFAC R.3.4). [8]

Column2: Transcription factor label (from TRANSFAC R.3.4). V means vertebrate. [8]

Column3: Similarity (0.0-1.0) between a registered sequence for the transcription factor binding sites and the input sequence (at the position shown in the next column). [8]

Column4: Position on the input sequence. [8]

Column5: Strandness. + and - means forward and reverse strands that the transcription factor binds, respectively. [8]

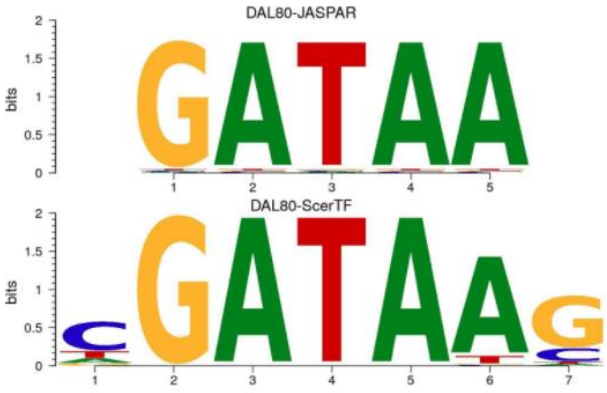
Column6: Consensus sequence (fixed) of the transcription factor binding sites. S = C or G, W = A or T, R = A or G, Y = C or T, K = G or T, M = A or C, N = any base pair. [8]

Column7: Subsequence from the input sequence at the position - corresponding to the consensus sequence. [8]

So, based on these results, all possible transcription factor binding sites are detected and they're ordered based on the matching scores. These results are all represented with a cut-off value specific for each entered sequence. In other words, this tool is unique in identifying the cut-off value for each entered transcription factor. This cut-off value allows the tool to separate the hits from non-binding sites. In general, the optimal cut-off value will accurately differentiate functional sites from background sequences. However, functional sites are not explicitly provided. Bucher's method extracts the transcription factor motifs and determines optimum cut-offs by maximizing the signal-to - noise ratio in the preferred region on the assumption that such functional sites are conserved in the region.

It didn't take much time to get the results, almost 3 seconds. This proves how functional and efficient is this tool till now since its publication.

For the other tool used, generegulation package, the results of searching for transcription factor binding sites for Dal80 were as follows:



An advantage of this package is that it can give results using 2 different databases: JASPAR and ScerTF based on the command line used as showed below:

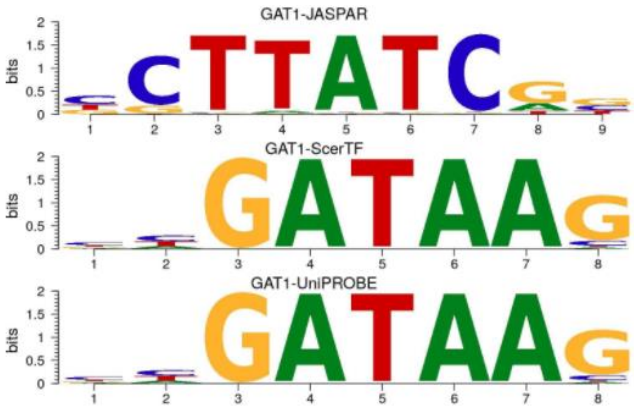
```
pfm.dal80.jaspar <- new("pfm", mat=query(MotifDb, "dal80")[[1]],
                        name="DAL80-JASPAR")
pfm.dal80.scertf <- new("pfm", mat=query(MotifDb, "dal80")[[2]],
                        name="DAL80-ScerTF")
plotMotifLogoStack(DNAmotifAlignment(c(pfm.dal80.scertf, pfm.dal80.jaspar)))
```

Among these two, the JASPAR motif has more data, but the ScerTF motif has been released more recently. ScerTF has a reputation for careful yeast-specific healing so it's preferred more for precise results.

It was proved that DAL80 "competes with Gat1 for binding suggesting that they would have highly similar motifs. That's why Gat1 was a good choice for getting results of to be compared with Dal80 for further investigation in this area. For getting the transcription factor binding motif of Gat1, the following command was used:

```
pfm.gat1.jaspar = new("pfm", mat=query(MotifDb, "gat1")[[1]],
                      name="GAT1-JASPAR")
pfm.gat1.scertf = new("pfm", mat=query(MotifDb, "gat1")[[2]],
                      name="GAT1-ScerTF")
pfm.gat1.uniprobe = new("pfm", mat=query(MotifDb, "gat1")[[3]],
                      name="GAT1-UniPROBE")
plotMotifLogoStack(c(pfm.gat1.uniprobe, pfm.gat1.scertf, pfm.gat1.jaspar))
```

That command searches 3 different databases: HASPAR, ScerTF, and UniPROBE. The results of that search were as follows:



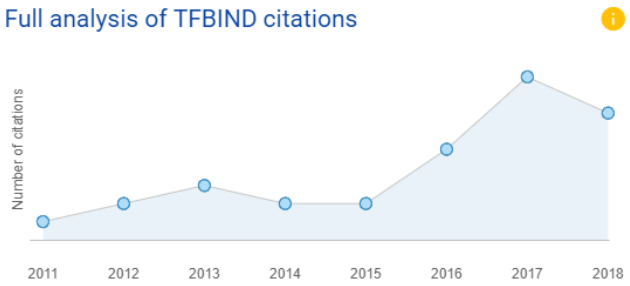
Based on these results, the GAT1-JASPAR motif is very similar to DAL80's GATAA motif, and thus consistent with the claim that GAT1 and DAL80 compete for binding. After getting these results, more work is needed to retrieve the promoter regions for a set of candidate targets, and identify the sequence matches of the binding motif in the genes' promoter regions. In this package, for the position weight matrix, they use a 90% minimum score for matching. It gave the following results:

```
## Views on a 1000-letter DNAString subject
## subject: TTGAGGAGTTGTCCACATACACATTAGTGTTGAT...GCAAAAAAAAAGTGAAATACTGCGAAGAAC
AAAG
## views:
##       start end width
##   [1]   620 626     7 [TGATAAG]
##   [2]   637 643     7 [CGATAAG]
```

After all, the online tool, TFBIND, can be considered a better one than the R package mentioned here because it gives all possible binding motifs for the entered transcription factor. On the other hand, the generegulation package works on getting the binding site from known databases and starts looking for highest matching scores with promoter regions. However, the generegulation package can be better in other investigations like finding or proving that Dal80 competes with Gat1 as mentioned earlier. That can be proved by comparing the binding motifs of both of them.

6. Conclusion and Discussion

To conclude, the TFBIND tool proved its precision in finding possible binding sites for entered transcription factors. By working on the 2 transcription factors, Dal 80 and GAT1, the tool gave an output for different binding sited ordered based on their matching scores. These matching scores were calculated after getting through the tool algorithm. The algorithm starts by getting the position frequency matrix and then it transferred it to a position weight matrix as illustrated previously. After all, the results were compared with the generegulation R package and differences were mentioned. The TFBIND tool proved its efficiency by how common it's used nowadays as shown below: [14]



6. Future work

The online tool, TFBIND, may need improvement in the visualization part for possible hits. That part can help in comparison of binding motifs of different transcription factors. Also, it would be better if they added another part for the tool to compare 2 transcription factors together and find if they both have similar motifs to indicate whether there is a competition binding or not.

8. References

[1]"Transcription: an overview of DNA transcription (article) | Khan Academy", Khan Academy, 2020. [Online]. Available: https://www.khanacademy.org/science/biology/gene-expression-central-dogma/transcription-of-dna-into-rna/a/overview-of-transcription.

[2]"transcription / DNA transcription | Learn Science at Scitable", Nature.com, 2020. [Online]. Available: https://www.nature.com/scitable/definition/transcription-dna-transcription-87/.

[3]"general transcription factor / transcription factor | Learn Science at Scitable", Nature.com, 2020. [Online]. Available: https://www.nature.com/scitable/definition/transcription-factor-167/.

[4]J. Tate, E. Tolley and T. Cooper, "Sit4 and PP2A Dephosphorylate Nitrogen Catabolite Repression-Sensitive Gln3 When TorC1 Is Up- as Well as Downregulated", Genetics, vol. 212, no. 4, pp. 1205-1225, 2019. Available: 10.1534/genetics.119.302371

[5]J. Crespo, T. Powers, B. Fowler and M. Hall, "The TOR-controlled transcription activators GLN3, RTG1, and RTG3 are regulated in response to intracellular levels of glutamine", Proceedings of the National Academy of Sciences, vol. 99, no. 10, pp. 6784-6789, 2002. Available: 10.1073/pnas.102687599

[6]"DAL80 | SGD", Yeastgenome.org, 2020. [Online]. Available: https://www.yeastgenome.org/locus/S000001742.

[7]"generegulation", Bioconductor, 2020. [Online]. Available: https://master.bioconductor.org/packages/release/workflows/html/generegulation.html.

[8]"TFBIND INPUT", Tfbind.hgc.jp, 2020. [Online]. Available: http://tfbind.hgc.jp/.

[9]"Position weight matrix - Dave Tang's blog", Dave Tang's blog, 2020. [Online]. Available: https://davetang.org/muse/2013/10/01/position-weight-matrix/.

[10]P. Bucher, "Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences", Journal of Molecular Biology, vol. 212, no. 4, pp. 563-578, 1990. Available: 10.1016/0022-2836(90)90223-9

[11]W. Wasserman and A. Sandelin, "Applied bioinformatics for the identification of regulatory elements", Nature Reviews Genetics, vol. 5, no. 4, pp. 276-287, 2004. Available: 10.1038/nrg1315

[12]"Saccharomyces cerevisiae S288C chromosome XI, complete sequence - Nucleotide - NCBI", Ncbi.nlm.nih.gov, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/nuccore/NC_001143.9?report=fasta&from=506898&to=507707.

[13]"Saccharomyces cerevisiae S288C chromosome VI, complete sequence - Nucleotide - NCBI", Ncbi.nlm.nih.gov, 2020.

[Online]. Available: https://www.ncbi.nlm.nih.gov/nuccore/NC_001138.5?report=fasta&from=95966&to=97498.

[14]"TFBIND analytics - omicX", Omictools.com, 2020. [Online]. Available: https://omictools.com/analytics/bioinformatics/software/tfbind.