IDENTIFICATION OF TRANSCRIPTION FACTOR BINDING SITES IN PROMOTER DATABASES

İlknur Melis Durası [1,*], *Uğur Dağ* [2,*], *Burcu Bakır Güngör* [1, 3], *Burcu Erdoğan* [2], *Işıl Aksan Kurnaz* [2] *and O. Uğur Sezerman* [1]

[1] Department of Biological Sciences and Bioengineering
Sabancı University, İstanbul, Turkey
ugur@sabanciuniv.edu; melisdurasi@sabanciuniv.edu

[2] Department of Genetics and Bioengineering
Yeidtepe University, İstanbul, Turkey
iakurnaz@yeditepe.edu.tr; ugur.dag@std.yeditepe.edu.tr; burcu.erdogan@std.yeditepe.edu.tr

[3] Department of Computer Science and Engineering
Bahçeşehir University, İstanbul, Turkey
burcu.gungor@bahcesehir.edu.tr

ABSTRACT

Transcription factors (TFs) are the proteins which regulates the expression of their target genes either in a positive or negative manner. TFs realize this task by binding to a specific DNA sequence contained in promoter regions, via their DNA binding motifs. Among ETS family TFs, Pea3 proteins are involved in the regulation of expression of genes, which are important for cell growth, development, differentiation, oncogenic transformation and apoptosis. *In silico* studies should be done to find out the novel target genes for this TF. Even though a few bioinformatics tools are available for this purpose, the user needs to go back and forth between different tools, and to repeat these steps for each of their candidate gene. Here we combined these tools and constituted a new tool which examines the affinity of any TF towards the selected target genes' promoter sequences. The tool is tested on several genes, which are predicted to be regulated by Pea3 TF.

I. INTRODUCTION

Transcription factors (TFs) are important proteins as they have key roles in regulation of their target gene's expression. They bind to the specific sequences on the genome which are called promoters. In certain cases, binding of multiple TFs are required to regulate the expression of the target gene. TFs can promote expression of a gene under certain physiological conditions and at specific cellular locations as well. It is crucial to determine the DNA sequences that TFs recognize, in order to find out which genes can be controlled by them. One way to determine these sequences is to find common DNA sequences in the upstream regions of the genes showing highly correlated expression level, assuming they are controlled by the same TF [1].

ETS (E26 transcription-specific) family of TFs are classified in the winged helix-turn-helix superfamily (wHTH) and posses a functional domain which involves evolutionarily preserved 85 amino acid residues. This conserved functional domain enables binding to a purine-rich DNA sequence with central 5'-GGAA/T-3' core sequence; DNA binding inhibition and transactivation. ETS family TFs can be subdivided into 30 protein members and one of its members is the Pea3 group proteins. Pea3 TFs are involved in the regulation of gene expression, which is important for cell growth, development, differentiation, oncogenic transformation and apoptosis. The subdivision of this family is done according to their distinct DNA binding specificities. The affinity of binding is determined by the ETS domain depending on whether it is in the amino terminal or carboxy terminal, the sequence of ETS domain and the DNA sequence neighbouring to the 5'-GGAA/T-3' central core [2].

Our preliminary studies focus on the identification of the downstream elements of Pea3 TF. In order to reveal novel target genes for Pea3, *in silico* studies should be performed as the first step. To realize this task, we first utilized online bioinformatics tools i.e., TFSearch, Transcriptional regulatory element database (TRED) [3] and Promo [4] manually, which requires extra time and it is an extremely cumbersome process. Next, to automate this process, we have developed a tool which utilizes existing bioinformatics tools and determines specific target genes and their promoter sequences and then examines specifically the affinity of the queried TF (Pea3 in this case) towards these sites.

In our new tool, specified regions (mostly upstream regions) of more than one gene of interest can be retrieved and automatically searched for binding motifs of queried TF.

*To whom correspondence should be addressed.

II. RELATED WORK

In literature, there are many different databases and tools to analyze and predict the promoter regions and binding sites of TFs. The discussion of all these tools is beyond the scope of this study. Instead, here we will review two most commonly used databases relevant to our goal.

First database of interest is TRED [3], which uses the cis- and trans- regulatory elements to provide access to the queried data by the user. The functionalities of TRED are: genome-wide mouse, human and rat promoter sequence annotation, TF binding and regulation information, sequence analysis tools, capability of retrieval of TF motifs, finding the relation of promoter sequences and TF binding information.

Second relevant database is Promo [4], which gives the information about the binding affinity of the selected TF in a single sequence or multiple sequences. The functionalities of Promo are: enabling selection of the related species or a group of species to retrieve the matrices that are related to the selected taxonomic level, giving information about the possible related genes that might be regulated by the queried TF, ability to analyze more than one sequence at a time. Promo uses the TRANSFAC database [6], which has the largest eukaryotic DNA binding sequences [4]. Hence, it is more advantageous to use Promo [4], compared to other publicly available databases [7].

III. METHOD

The proposed work aims to gather information from two separate databases and combining their results in an automated way.

The user interface of the tool developed allows the user to:

i. select the species among human, mouse or rat;

ii. select or enter the gene names that the user is interested in and the region of the sequence where the TF motifs may be found;

iii. select the TFs that are wanted to be searched on the selected potential promoter sequences.

In the first part of this study, TRED database is used to get the user specified regions (mostly upstream regions) of the user specified gene(s) where the promoter sequences may be found, as shown in Figure 1. As a result of this step, our tool gets the promoter sequences in FASTA format. In case the promoter sequences will be used for further analysis, it downloads the information about potential promoter sequence regions to a file called *FastaSeq_Result.txt* as shown in Figure 2.

In order to search for the binding regions of the queried TF (Pea3 in this study) on these potential promoter sequences, the second database 'Promo' is used, which constitutes the second part of our tool. Even though the first part receives the specie(s) input through the interface, the second part of the tool does not ask the user to select any species because the default selection is already set to '*All factors*' -'*All sites*'. Next, our tool wants the user to select the TF that is wanted to be searched. The tool also gives the option of searching multiple TFs at the same time. For the purpose of this study, the TF Pea3 is chosen. After the selection of the TF, the program automatically submits the promoter sequences that were in FASTA format to be able to get the final analysis results from Promo. This step gets the profile matrices of the chosen TF and calculates the binding affinities against the submitted promoter sequences. The final result of the analysis is given to the user in the file named *TF_Result.txt*. In Figure 3, the format of the final result can be seen. For each promoter sequence that is analysed, there is information about the sequence name, TF's name, its start and end position, dissimilarity value -that we are mostly interested in, the string and the random expectation (RE) values i.e. RE equally and RE query [5].

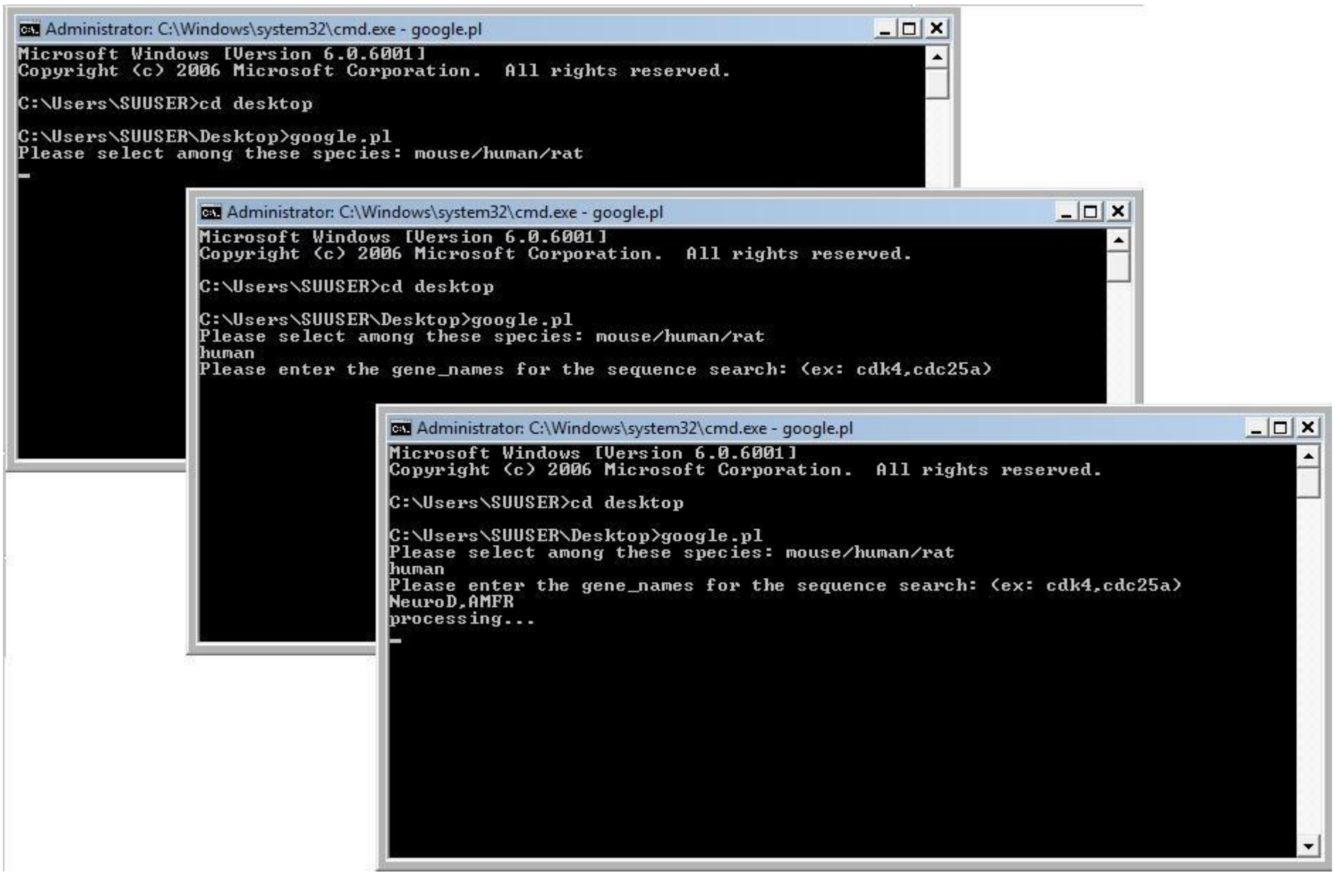
The dissimilarity rate measures the variation of the found TF binding sequence from the known motifs for the queried TF. The tool searches for motifs with the dissimilarity margin less than or equal to %15, which is selected as the threshold. Among all of the final dissimilarity values, the ones with the minimum dissimilarity rates are pointed out to be used for further studies.

The sequences that are taken from TRED in the first step can also be used for scanning for multiple TF binding motifs automatically. Our tool is developed in Perl and the source code is available upon request.

IV. RESULTS

In this study, since we are mainly focused on Pea3 TF, the genes that might be related to the Pea3 TF are used for the analysis. We initiated our search to query the Pea3 binding sites of NeuroD and AMFR (Autocrine Motility Factor Receptor) genes. These genes are functionally related and assumed to be controlled by Pea3.

AMFR is recognized as metastatic gene marker for breast cancer [8]. Since Pea3 is a pivotal regulator for breast cancer progression [9], AMFR is considered as a good target for Pea3. By using our newly developed tool, Table 1 lists the binding affinity scores of Pea3 against several candidate genes' promoters. AMFR has the dissimilarity score of 0, showing that Pea3 motif occurs in the upstream region of this gene and this gene is highly likely to be regulated by Pea3. We are currently in the process of cloning the AMFR promoter region into the expression vector to verify the binding of Pea3 to this region.


Administrator: C:\Windows\system32\cmd.exe - google.pl

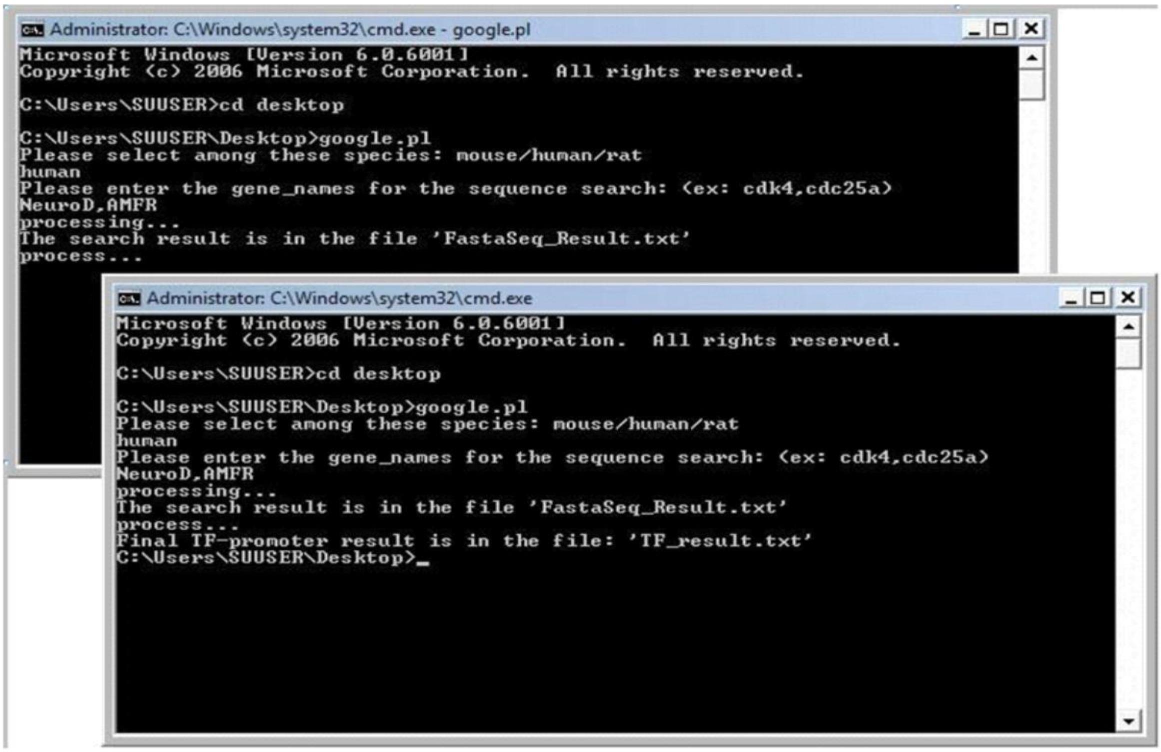
```
Microsoft Windows [Version 6.0.6001]
Copyright (c) 2006 Microsoft Corporation.  All rights reserved.

C:\Users\SUUSER>cd desktop

C:\Users\SUUSER\Desktop>google.pl
Please select among these species: mouse/human/rat
_
```

Administrator: C:\Windows\system32\cmd.exe - google.pl

```
Microsoft Windows [Version 6.0.6001]
Copyright (c) 2006 Microsoft Corporation.  All rights reserved.

C:\Users\SUUSER>cd desktop

C:\Users\SUUSER\Desktop>google.pl
Please select among these species: mouse/human/rat
human
Please enter the gene_names for the sequence search: (ex: cdk4,cdc25a)
```

```
Microsoft Windows [Version 6.0.6001]
Copyright (c) 2006 Microsoft Corporation.  All rights reserved.

C:\Users\SUUSER>cd desktop

C:\Users\SUUSER\Desktop>google.pl
Please select among these species: mouse/human/rat
human
Please enter the gene_names for the sequence search: (ex: cdk4,cdc25a)
NeuroD,AMFR
processing...
_
```

Figure 1. The flow and the user interface part of the tool.

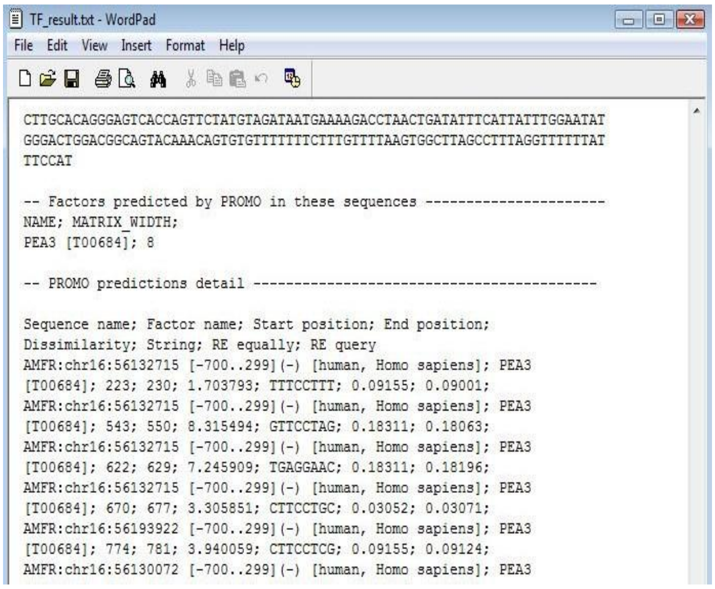

Administrator: C:\Windows\system32\cmd.exe - google.pl

```
Microsoft Windows [Version 6.0.6001]
Copyright (c) 2006 Microsoft Corporation.  All rights reserved.

C:\Users\SUUSER>cd desktop

C:\Users\SUUSER\Desktop>google.pl
Please select among these species: mouse/human/rat
human
Please enter the gene_names for the sequence search: (ex: cdk4,cdc25a)
NeuroD,AMFR
processing...
The search result is in the file 'FastaSeq_Result.txt'
process...
```

Administrator: C:\Windows\system32\cmd.exe

```
Microsoft Windows [Version 6.0.6001]
Copyright (c) 2006 Microsoft Corporation.  All rights reserved.

C:\Users\SUUSER>cd desktop

C:\Users\SUUSER\Desktop>google.pl
Please select among these species: mouse/human/rat
human
Please enter the gene_names for the sequence search: (ex: cdk4,cdc25a)
NeuroD,AMFR
processing...
The search result is in the file 'FastaSeq_Result.txt'
process...
Final TF-promoter result is in the file: 'TF_result.txt'
C:\Users\SUUSER\Desktop>_
```

Figure 2. The execution of the program, which generates two output files: the sequence and the TF-promoter sequence relational result.



```
CTTGCACAGGGAGTCACCAGTTCTATGTAGATAATGAAAAGACCTAACTGATATTTCATTATTTGGAATAT
GGGACTGGACGGCAGTACAAACAGTGTGTTTTTTTCTTTGTTTTAAGTGGCTTAGCCTTTAGGTTTTTTAT
TTCCAT

-- Factors predicted by PROMO in these sequences ----------------------
NAME; MATRIX_WIDTH;
PEA3 [T00684]; 8

-- PROMO predictions detail -----------------------------------------

Sequence name; Factor name; Start position; End position;
Dissimilarity; String; RE equally; RE query
AMFR:chr16:56132715 [-700..299](-) [human, Homo sapiens]; PEA3
[T00684]; 223; 230; 1.703793; TTTCCTTT; 0.09155; 0.09001;
AMFR:chr16:56132715 [-700..299](-) [human, Homo sapiens]; PEA3
[T00684]; 543; 550; 8.315494; GTTCCTAG; 0.18311; 0.18063;
AMFR:chr16:56132715 [-700..299](-) [human, Homo sapiens]; PEA3
[T00684]; 622; 629; 7.245909; TGAGGAAC; 0.18311; 0.18196;
AMFR:chr16:56132715 [-700..299](-) [human, Homo sapiens]; PEA3
[T00684]; 670; 677; 3.305851; CTTCCTGC; 0.03052; 0.03071;
AMFR:chr16:56193922 [-700..299](-) [human, Homo sapiens]; PEA3
[T00684]; 774; 781; 3.940059; CTTCCTCG; 0.09155; 0.09124;
AMFR:chr16:56130072 [-700..299](-) [human, Homo sapiens]; PEA3
```

Figure 3. The final result in txt format presents all the prediction details for predictions with dissimilarity rate <= 15.

The involvement of NeuroD in neuronal differentiation [5] as our TF makes NeuroD a putative target for Pea3. This putative regulatory relation between NeuroD and Pea3 has been also shown by luciferase reporter assays; such that the presence of Pea3 significantly up-regulated the NeuroD promoter. Binding affinity of Pea3 to MMP-9 (matrix metalloprotease 9) has been previously shown in literature [10], thus MMP-9 is used as positive control. As expected, we observed that the binding affinity of Pea3 to the promoter region of MMP-9 is relatively high (0.0% dissimilarity rate for Promo 3.0)

In addition to those, the genes taking roles in nervous system development, i.e, Ephrin-B2 and Presenilin-1; in extracellular matrix organization, i.e, RECK (Reversion-Inducing-Cysteine-Rich- Protein), Alox15b (Arachidonate 15-lipoxygenase type B) and in angiogenesis i.e, angiopoietin were scanned to reveal the possible TF binding sites within their promoter regions. As mentioned earlier, the cellular functions of these genes overlap with the downstream effects of the Pea3 TF, thus making them considerable for being targets of Pea3. The binding affinity of Pea3 to the promoter regions of these genes of interest is also summarized in Table 1, using the minimum dissimilarity rate measure.

V. CONCLUSION

In this study, we have analysed the promoter regions of the genes of interest in terms of the TF Pea3 binding affinity. Pea3 is found to have key roles in regulation of gene expression, related to cell growth, oncogenic transformations and metastasis, NeuroD and AMFR are selected among many other genes as they are functionally similar to each other and they both may be regulated by Pea3. The analysis shows that both have Pea3 binding sites. The new tool that has been developed uses online available tools TRED and Promo database to automatically determine the binding affinities of the upstream re regions of these genes to Pea3. In future studies, we aim to upgrade our tool So that it will be possible to investigate all the genes for the selected organism and display the binding affinities of the user selected TFs to these genes ranked according to their binding affinity. Hence, querying via TFs name, novel promoter regions could be discovered in shorter time frame. The new tool will search the whole genome for the possible binding sites of the selected TF. Users will also be able to search for genes regulated by multiple TFs. The tool will give the user, the related gene list possibly regulated by the selected transcription factors both individually and together.

We hope our study facilitates the research in this field via automatically detecting the binding affinities of selected TFs to their target genes. This could be especially important for the TFs such as Pea3, which is used here as a case study, to prevent neurological disorders and to find solutions to cancer patients in the risk of metastasis.

Gene	Source	Process	Pea3 Binding Property(minimum dissimilarity percentage rate)
NSC1 – Nonselective cation channel 1	*Mus musculus*	Methyltransferase Activity	1.70
Alox15b - Arachidonate 15-lipoxygenase, type B	*Mus musculus*	Regulation of cell proliferation and migration	0.21
Hfe - Hemochromatosis	*Homo sapiens*	Antigen Processing And Presentation	1.07
RECK - Reversion-Inducing-Cysteine-Rich Protein	*Homo sapiens*	Extracellular Matrix Organization	0.63
ANGPT – 1 Angiopoietin	*Homo sapiens*	Angiogenesis	0.21
C4.4A – Metastasis associated GPI-anchored protein	*Rattus norvegicus*	Cell matrix adhesion	0.21
PSEN 1 – Presenilin – 1	*Homo sapiens*	Epithelial cell proliferation and brain development	0.21
UMOD – Uromodulin	*Homo sapiens*	Regulation of cell proliferation	0
MMP7 – Matrilysin	*Mus musculus*	Degradation of ECM	0.63
CTNNA3 – alphaT-catenin	*Homo sapiens*	Cell-cell adhesion	0.21
EFNB2 – ephrin-B2,	*Homo sapiens*	Nervous system development	0
AMFR – autocrine motility factor receptor	*Homo sapiens*	Signal transduction for metastasis	0

Table 1. The genes that are used in the process of testing the tool and the dissimilarity scores of the potential promoter regions.

REFERENCES

[1] Veerla S., Ringner M., Höglund M., "Genome-wide transcription factor binding site/promoter databases for the analysis of gene sets and co-occurence of transcription factor binding motifs", *BMC Genomics*., **11**:145, 2010

[2] Graves BJ, Petersen JM, "Specificity within the ETS family of transcription factors", *Advences in Cancer Research* **75**:1-55., 1998

[3] Zhao F., Xuan Z., Liu L., Zhang M.Q., "TRED: a Transcriptional Regulatory Element Database and a platform for *in silico* gene regulation studies", *Nucleic Acids Research,* vol.33, D103-D107, 2005

[4] Farre D., Roset R., Huerta M., Adsuara J.E., Rosello L., Alba M.M., Messeguer X., "Identification of patterns in biological sequences at the ALGGEN server: PROMO and MALGEN", *Nucleic Acids Reasearch*, vol.31, pp. 3651-3653, 2003

[5] Schwab, M. H., Bartholomae, A., Heimrich, B., Feldmeyer, D., Druffel-Augustin, S., Goebbels, S., Naya, F. J., Zhao, S., Frotscher, M., Tsai, M. J., and Nave, K. A., "Neuronal basic helix-loop-helix proteins (NEX and BETA2/Neuro D) regulate terminal granule cell differentiation in the hippocampus," *J. Neurosci.*, 3714–3724., 2000

[6]Matys V., Kel-Margoulis O.V., Fricke E., Liebich I., Land S., Barre-Dirrie A., Reuter I., Chekmenev D., Krull M., Hornischer K., Voss N., Stegmaier P., Lewicki-Potapov B., Saxel H., Kel A.E., Wingender E., "TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes", *Nucleic Acids Research*, vol. 34, D108-D110, 2006

[7]Heinemeyer T., Wingender E., Reuter I., Hermjakob H., Kel A.E., Kel O.V., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Kolpakov F.A., Podkolodny N.L., Kolchanov N.A., "Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL", *Nucleic Acids Research*, vol. 26, pp. 362-367. , 1998

[8] Baumann M., Kappl A., Lang T., Brand K., Siegfried W. & Pa E., "The diagnostic validity of the serum tumor marker phosphohexose isomerase (PHI) in patients with gastrointestinal, kidney, and breast cancer." *Cancer Investigation* **8**:351–356. , 1990

[9] Launoit Y., Baert J., Lelievre AC, *et.al*, "The Ets Transcription Factors of the PEA3 Group: Transcriptional Regulators in Metastasis", *Biochemica et Biophysica Acta* **1766**:79-87. , 2006

[10] Dahl KDC. , Zeineldin R. and Hudson LG. , " Pea3 is necessary for optimal growth factor receptor-stimulated matrix metalloproteinase expression and invasion of ovarian tumor cells" , *Mol. Cancer Res.* 5: 413-421, 2007.