

cisAnalyzer and PASPdb are novel tools for sequences analyses

The study of plant transcriptional initiation regions is a complex issue as multiple elements get involved in this level of regulation (Brkljacic and Grotewold, 2017). Additionally, an important lack of specialized databases when compared to model species (Yilmaz *et al.*, 2011; Chow *et al.*, 2016) and reports about experimentally proven peach transcription factors and binding sites was specially evidenced for *P. persica* [L.] Batsch, agreeing with the minor knowledge of Rosaceae fruits-specific mechanisms of expression regulation (Farinati *et al.*, 2017). In order to contribute to transcriptional initiation analyses independently from species, to add analysis to classical results and to fasten their processing and interpretation, a novel bioinformatic tool for cis-elements analyses, and a new plant TF binding sites database were constructed.

cisAnalyzer: a novel tool for motif search, enrichment, localization, co-occurrence and discovery

A great diversity of programs and on-line resources were created to mainly identify cis-elements, utilizing DNA sequences as input (Hertz and Stormo, 1999; Thompson *et al.*, 2003; Peng *et al.*, 2006; Bailey *et al.*, 2015; Chow *et al.*, 2016). However, we designed a novel program to contribute to transcriptional initiation analyses independently of species and to adding analysis to classical motif searches results. cisAnalyzer is a simple but versatile program, Unix/Linux-implemented (console application) and based on Perl code and with graphical output generation through R calls. The user's input was considered as the main control mechanism which leads the type of analysis to be done. The format of sequences input files was also taken into account leaving the features of Biomart (Phytozome) multifasta output files as the ones needed for the program to work. PASPdb was included in cisAnalyzer as one of three options of plant DNA (regulatory sequences). Other options included custom (DNA, RNA and protein) and unknown motif (DNA and RNA) analyses, augmenting the program capabilities and user implementations. The Unix/Linux-based implementation of cisAnalyzer maybe an initial limitation for several users but we recommend it because of the capabilities of these or related operating systems to process large files of sequence data, bringing this often needed type of searches closer to wet lab scientists. Several steps of target sequence(s) and motif(s) quality control and handling were included along with the ulterior steps of motif search, enrichment, localization and co-occurrence analyses, and report of the found results. The classic motif results tables, novel tabular reports with sequence-specific information, and R graphical outputs are generated. These advances accelerate the ulterior processing and interpretation by providing to the user a great diversity of motif and target sequences data to better interpret the results and conclude about their biological significance. cisAnalyzer and all its program features are open to be improved for the benefit of its users.

PASPdb: Plant Abiotic Stress- and Phytohormone signaling-related cis-elements database

Based on the acceptance of searching known motifs from different plant species when analysing promoters as DNA binding properties of TF families are conserved (Wisniewski *et al.*, 2006; Artlip *et al.*, 2013; Pons *et al.*, 2014; Zhou *et al.*, 2014; Genero *et al.*, 2016; Wang *et al.*, 2017), we created PASPdb. We exhaustively gathered plant TF-related cis-elements from motif databases footprintDB (Sebastian and Contreras-Moreira, 2014) and PlantPAN (Chow *et al.*, 2016), which comprises other classical databases of plant motifs as PLACE, AGRIS, TRANSFAC, JASPAR, and PlantCARE, more than 300 experimentally proven cis-elements and their descriptions were collected for a novel database creation. The selection was based in associated-literature verification and the common features of the selected motifs were the existence of their in vivo and/or in vitro experimentally proven functionality. Moreover, aiming to fasten the ulterior sub-analysis of detected cis-elements, the reported functionalities of the selected motifs also allowed manually grouping them in clusters based on the phytohormone, abiotic or biotic stimulus under which they have a biological role. The user could now restrict its analyses selecting particular group(s) of related motifs and could rapidly interpret potential roles of the identified cis-elements. 18 groups were constructed, including related to: light, cold, heat, oxidative, wounding, dehydration (abscisic acid (ABA)-dependent), dehydration (ABA-independent), Unfolded Protein Response (UPR), sugar response, circadian rhythm, auxin (AUX), cytokinin (CK), ethylene (Et), jasmonic acid (JA), salicylic acid (SA), gibberellins (GA) and a miscellaneous group. A particular consideration is the literature-driven complexity of the performed grouping due to the existence of TF family members that potentially interact with the same motifs but possess different sensitivities to different stimuli and even of certain TFs with the capacity of binding to more than one cis-element (Sun *et al.*, 2008; Cheng *et al.*, 2013). PASPdb curated list of plant TF-binding sites from multiple species and each motif-associated reference is provided in Supplementary Table S13 and also included in /cisAnalyzer/db/ folder along with particular PASPdb files allowing its usage from the software. New motifs could be integrated and new groups could be created for users allowing PASPdb to grow.

Previous requirements

Unix/Linux or Unix/Linux-related operating system

Perl

There is probably one Perl interpreter already installed by default in your Unix/Linux operating system. You could get Perl at <https://www.perl.org/get.html>. You could verify the installed Perl version with:

```
:~$ perl -v
```

Perl Modules List::MoreUtils, CGI and GD::Graph

You could start CPAN (Comprehensive Perl Archive Network, <https://www.cpan.org/>) shell and install both needed modules running the commands indicated in <http://www.walkingrandomly.com/?p=55> or <http://www.tldp.org/LDP/LG/issue81/padala.html>.

R

You could install R in your Unix/Linux operating system, following the instructions from <https://cran.r-project.org/>

R packages gplots and RcolorBrewer

You could install them easily from R. Problems with these R packages could be evidenced at /MyOutputFiles folder if there are no generated graphs. Additionally, searching for errors during R system call in .Rout file at /cisAnalyzer/tmp folder should help for solving the issue.

cisAnalyzer installation and contents

Supplementary compressed file cisAnalyzer.zip contains all requirements for its installation and use, including the source codes and developed scripts and modules which are freely available. As many other Perl scripts, there is no need of specific installation procedures. The user must only unzip the zip file in a desired path. As an example, you could copy to and unzip cisAnalyzer from terminal in a Perl folder included in Desktop folder:

copying the zip file to a desired folder

```
:~$ cp Downloads/cisAnalyzer.zip /Desktop/Perl/cisAnalyzer.zip
```

changing to the desired location

```
:~$ cd Desktop/Perl/
```

unzipping cisAnalyzer folder

```
:~/Desktop/Perl$ unzip cisAnalyzer.zip
```

The cisAnalyzer structure of folders and files will appear at a new cisAnalyzer uncompressed folder and the user should navigate inside for using:

```
:~/Desktop/Perl$ cd cisAnalyzer/
```

```
:~/Desktop/Perl/cisAnalyzer$
```

cisAnalyzer unzipped folder comprises the following PASFdb, test, input and output folders:

```
~/Desktop/Perl/cisAnalyzer/Test Files/Test targets.txt
~/Desktop/Perl/cisAnalyzer/MyMotifs
~/Desktop/Perl/cisAnalyzer/db/PASFdb Perl.csv
~/Desktop/Perl/cisAnalyzer/db/PASFdb FULL Perl.csv
~/Desktop/Perl/cisAnalyzer/db/PASFdb lettersandgroups Perl.csv
~/Desktop/Perl/cisAnalyzer/MyTargets
~/Desktop/Perl/cisAnalyzer/MyTarget3UTRs
~/Desktop/Perl/cisAnalyzer/MyTarget5UTRs
~/Desktop/Perl/cisAnalyzer/tmp/DNA/ with 18 .R files
~/Desktop/Perl/cisAnalyzer/tmp/RNA/ with 18 .R files
~/Desktop/Perl/cisAnalyzer/tmp/protein/ with 9 .R files
~/Desktop/Perl/cisAnalyzer/MyOutputFiles
```

cisAnalyzer unzipped folder comprises the following script, modules and tutorial files:

```
~/Desktop/Perl/cisAnalyzer/Start_cisAnalyzer.pl
~/Desktop/Perl/cisAnalyzer/Run_cisAnalyzer.pl
~/Desktop/Perl/cisAnalyzer/cisAnalyzer_modules.pm
~/Desktop/Perl/cisAnalyzer/cisAnalyzer_modules2.pm
~/Desktop/Perl/cisAnalyzer/cisAnalyzer_tutorial.pdf
```

cisAnalyzer capabilities

cisAnalyzer and PASPdb were originally developed to assist *P. persica* [L.] Batsch promoter and regulatory sequences studies but related and other variations of were included, widening its uses. Table 1 summarizes main features of each type of analysis.

Table 1. cisAnalyzer capabilities of analysis.

Option	Target sequence(s)		Motif(s)	
	Type	Origin	Type	Origin
Custom motif analysis	Nucleotidic Aminoacidic	User	Nucleotidic Aminoacidic	User
Unknown motif enrichment analysis	Nucleotidic	User	Nucleotidic	cisAnalyzer detects enriched motifs
PASPdb-assisted known motif analysis	Plant DNA regulatory sequences	User	DNA	PASPdb

As it could be seen from table 1, cisAnalyzer offers three main options of analysis. A CUSTOM MOTIF analysis (C) allows the user to work with DNA, RNA or aminoacidic sequences providing targets to be searched and motifs to search. This allows potentially a great number of analysis (e.g. user new discovered or PASPdb-not included TF binding sites in promoters, restriction enzymes sites, ribonucleic motifs for splicing factors or RNA binding proteins, post-translational modification o ligand binding sites or subcellular localization signals in polypeptides, etc...) and offers a tool for fast finding and reporting cis-elements pattern matchings in each type of biological sequences. On the other hand, a PASPdb-assisted KNOWN MOTIF (P) analysis allows the characterization of plant DNA regulatory regions related to transcriptional initiation with the plant collected TF binding sites of the novel database. A special advantage of PASPdb use is to fasten the analysis by selecting particular motif/s, particular group/s of related motifs, phytohormone signaling-groups, abiotic stress-groups, or all groups of related motifs. Additionally, the utilization of the curated database allows to potentially associate target sequences containing found motifs with particular stimuli of interest. Finally, performing an UNKNOWN MOTIF ENRICHMENT (U) analysis becomes useful for motif discovery in a set of nucleotidic related sequences as it finds lineal patterns whose presence is distributed (at least with single presence) in half or more targets.

cisAnalyzer general step-by-step use

Figure 1 summarizes main steps in cisAnalyzer functioning for the different options of analysis.

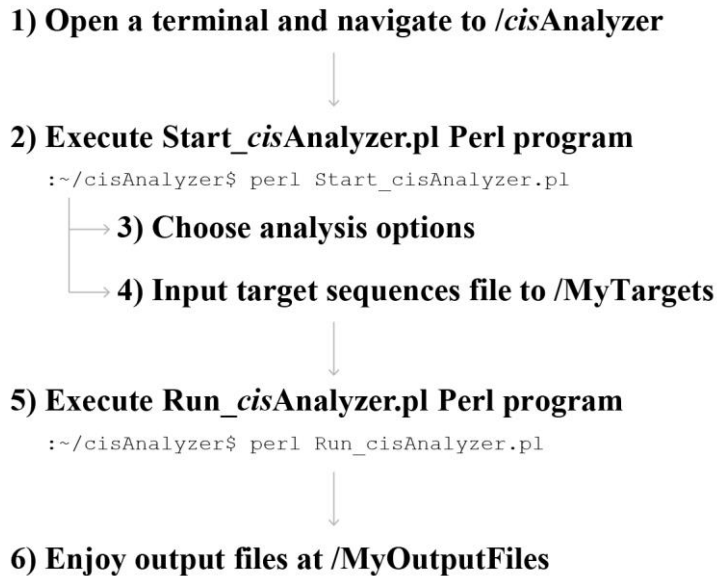


Figure 1. Overview of main cisAnalyzer functioning steps.

At cisAnalyzer folder and after Start_cisAnalyzer.pl execution, the user is guided step-by-step while making analysis options decisions and inputting needed files. Only for a first time, the user will be asked to enter the path to cisAnalyzer folder. After that, the program will identify the path automatically (from a generated path file at /cisAnalyzer folder). Then, the user will be asked to enter its name (that also could work as a denomination for the analysis). We recommend this name to be short and capable of identify the particular analysis. Next, the program receives user's analysis options (saved in a particular output file) and the file of target sequences (and could receive UTR(s) sequences and motif files), process, controls and informs (at /MyOutputFiles folder) sequences quality, and generates processed files for following steps. After this program running, a path file is generated at /cisAnalyzer as described while other automatically generated files at /cisAnalyzer/tmp folder are date, user, cisA_identifiers, cisA_Motifs, and cisA_target_sequences.

Inside CUSTOM MOTIF analysis (C), the user first faces the following options and suboptions that define the analysis and offer biological information for outputs generation:

- D) DNA
- R) RNA
- P) Protein

CUSTOM MOTIF analysis (C) comprises the use of user's own motifs. This enriches cisAnalyzer capabilities as it allows wet lab scientists fastening the obtention of results regarding their own discovered or interest cis-elements. To do this, the user must input a motif .txt file in /MyMotifs folder and ensure the format have the following format (example for DNA analysis):

```
>TNTTAATTTA
>TAATTCTT
>GGTSTGA
```

The sequences must be written without blanks and with IUPAC (International Union of Pure and Applied Chemistry) characters as it is often needed for transcription factors binding sites. Nucleotide IUPAC characters are: A, C, G, T, Y=C/T, R=A/G, W=A/T, S=G/C, M=A/C, K=T/G, V=A/C/G, B=C/G/T, D=A/G/T, H=A/C/T, N=A/C/G/T; and their equivalents with U for RNA sequences. Aminoacidic IUPAC characters are: A=Ala, C=Cys, G=Gly, H=His, I=Ile, P=Pro, S=Ser, T=Thr, V=Val, L=Leu, M=Met, F=Phe, W=Trp, Y=Tyr, N=Asn, Q=Gln, E=Glu, D=Asp, R=Arg, K=Lys, y X=Any. Sequences/motifs with incorrect characters (for the chosen options) will be removed as well as motifs whose length exceeds the length of at least one valid target sequence. cisAnalyzer will report the invalid motifs in particular files in /MyOutputFiles folder.

After that and inside D and R options, the user faces other following questions (P option goes on without any extra choice):

D) DNA

- 1) Any DNA sequences (including promoter sequences)
- 2) Promoter + 5'UTR sequences
- 3) Promoter without 5'UTR sequences

R) RNA

- 1) Any RNA sequences (including mRNA-coding sequences (without considering UTRs))
- 2) Full mRNA sequences (including both 5' and 3'UTR in consideration)

Including UTRs sequences (C-D-2 and C-R-2 options) allows showing cis-element information about these regions and could be very useful for biologists. But it must be noticed that UTR results are subjected to their annotation (the user could verify how many sequences have annotated UTRs). D, R and P options and suboptions are followed by an extra questioning about the similarity of the sequences lengths within the set of targets, to set different localization analyses features in output files:

- 1) Target sequences with the same lenght
- 2) Target sequences with the different lenght

An UNKNOWN MOTIF ENRICHMENT (U) analysis is a motif discovery approach. De novo cis-element finding is usually started joining related sequences and looking for the enriched motifs in those sets as certain similarities between biologically-related sequences (e.g. co-regulated promoters) could exist. The search cisAnalyzer offers is initiated with 5-nucleotide motifs randomly created from A, C, G and T nitrogenous bases and extended to longer motifs when its predecessor possesses 0.5 or higher enrichment frequency. Users could have a prominent participation in the discovery as they know and could fine tune the quality of the target sequences biological relationship, and also targets length. Initially, the user should choose the nature of the sequences to work with:

D) DNA

- 1) Any DNA sequences (including promoter sequences)
- 2) Promoter + 5'UTR sequences
- 3) Promoter without 5'UTR sequences

R) RNA

- 1) Any RNA sequences (including mRNA-coding sequences (without considering UTRs))
- 2) Full mRNA sequences (including both 5' and 3'UTR in consideration)

U-D, and U-R options and suboptions are followed by the question about the similarity of the target sequences lengths:

- 1) Target sequences with the same lenght
- 2) Target sequences with the different lenght

Plant promoters and regulatory regions could be analyzed with PASPdb-assisted KNOWN MOTIF (P) analysis. Initially, 5 options are presented to the user:

- 1) Selecting particular motifs from preformed group(s) of related cis-elements
- 2) Selecting entire preformed group(s) of related cis-elements
- 3) Selecting the motifs from all our preformed ABIOTIC STIMULI-related groups of cis-elements
- 4) Selecting the motifs from all our preformed PHYTOHORMONE SIGNALLING-related groups of cis-elements
- 5) Selecting the motifs from ALL our preformed groups of cis-elements

After P-1 or P-2 choices, the user should define how many and which motifs (P-1) or group of motifs (P-2) will be used for the analysis (this is achieved by typing the corresponding identifiers of groups and motifs when asked). P-3, P-4 and P-5 options work as fast choices for low information sequences

analysis. It must be noticed that other known cis-elements could be desired for analysis of plant promoters but they are not included in PASPdb, the user could search them through custom motif analysis. Selected PASPdb-related motifs and/or groups of motifs are reported in files during and after the user selection process. The user has to choose the nature of the sequences to work with:

- 1) Any DNA sequences (including promoter sequences)
- 2) Promoter + 5'UTR sequences
- 3) Promoter without 5'UTR sequences

All PASPdb-related options and suboptions are followed by the question about the similarity of the target sequences lengths:

- 1) Target sequences with the same length
- 2) Target sequences with the different length

Finally, a common feature for mentioned C, P and U options and suboptions of analysis is the input of a target sequences file, ensuring its unique presence in /MyTargets folder (at terminal, you could use `$rm -f *` command inside /MyTargets folder: <https://www.computerhope.com/unix/urm.htm>). Also, the user must ensure a Biomart (Phytozome v12.1, Goodstein *et al.*, 2012)-derived specific fasta format as the following example shows (for DNA analysis):

```
>GeneName|TranscriptName|Organism name
TTTTAATTTATAATTCTATTGGTTTCACGATTGGTTTGGGTGCCGAGG
>GeneName2|TranscriptName2|Organism name2
TAGTGTAGTACGATGTAGATTTTGTGGGTGTGTCCACACAGAGAGTATACATCATTACAGTC
```

It should be noted that the utilization of Phytozome (and others) genome data is subjected to its publication and restriction policies; the user is responsible for the given use of the sequences. Sequences with non-IUPAC characters and/or spaces will be removed but cisAnalyzer will report them in particular files in /MyOutputFiles folder. The same requirements are imposed when entering 5' and/or 3'UTR sequences files. Although cisAnalyzer allows the user working with sequences independently from species, only Phytozome v12.1 (plants and related organisms) genome denominations (July 2018) were included in Start_cisAnalyzer.pl program. Other plant species (and organism names declared in fasta headers) could be input and their denominations must be inserted when cisAnalyzer asks:

```
> Do all your target sequences belong to one or more species included in
Phytozome? (Y/N)
```

When at least one sequence to be analysed belongs to a non-Phytozome species, you should answer the above question with N. Then you insert the number of non-Phytozome species:

```
> please answer, how many species not included in Phytozome are within your  
set of sequences?:
```

And finally you should type in the denomination of the desired species (Species name 1, etc...). Not inserting the particular denominations of non-Phytozome species will cause the removal of this/these sequence(s) before Run_cisAnalyzer.pl execution and the user could not get the results of its/their desired analysis.

Start_cisAnalyzer.pl ends, its processing results could be obtained from /MyOutputFiles, and without any extra user input, **Run_cisAnalyzer.pl execution** comprises the suitable motif pattern matching analysis over the suitable target sequences. It executes the subroutines included in the novel modules cisAnalyzer_modules.pm and cisAnalyzer_modules2.pm. The designed pattern matching algorithm searches globally each motif in each target sequence and down-processes the matching results in order to obtain non-classical report files. Specific motif variants included in the consensus (as TATSTAT = TAT[G|C]TAT) are particularly searched (TATGTAT and TATCTAT) and their results saved but they are also combined for the consensus motif results. Positive pattern matching runs will generate the following files at /MyOutputFiles:

User's decisions-containing file

~/Desktop/Perl/cisAnalyzer/MyOutputFiles/cisA_user_date_Made decisions.txt

Start_cisAnalyzer.pl target sequences-processing results

~/Desktop/Perl/cisAnalyzer/MyOutputFiles/cisA_user_date_Input seqs.txt

~/Desktop/Perl/cisAnalyzer/MyOutputFiles/cisA_user_date_Removed seqs.txt

PASPdb pre-selected group(s) of motifs (when chosen)

~/Desktop/Perl/cisAnalyzer/MyOutputFiles/cisA_user_date_Selected PASPdb group(s).txt

PASPdb effectively selected motif(s) with their information

~/Desktop/Perl/cisAnalyzer/MyOutputFiles/cisA_user_date_Selected PASPdb motif(s).txt

Start_cisAnalyzer.pl custom motifs-processing results (when chosen)

~/Desktop/Perl/cisAnalyzer/MyOutputFiles/cisA_user_date_Input motifs.txt

~/Desktop/Perl/cisAnalyzer/MyOutputFiles/cisA_user_date_Removed seqs.txt

Start_cisAnalyzer.pl unknown motifs discovery results (when chosen)

~/Desktop/Perl/cisAnalyzer/MyOutputFiles/cisA_user_date_Motif discovery.txt

Custom cisA_user_date_Input motifs.txt and unknown motif discovery cisA_user_date_Motif discovery.txt files also show a comparison of the motifs with PAsPdb motifs, revealing the presence of (PAsPdb-included) known motifs within your own or discovered motifs set.

The obtained results are reported in classical tables with motif and target sequences id information. The user could browse particular results from this file.

~/Desktop/Perl/cisAnalyzer/MyOutputFiles/cisA_user_date_Table.xls

The novelty of cisAnalyzer resides in the down-processing of classical motif searchings results, and its results are shown in the following battery of output files, aiming to fasten the biological interpretation of cis-elements finding in sequences. To do this, the obtained results are part of two types of constructed relative frequencies: those relative to motif matches and those relative to the number of targets with at least one motif match. Moreover, 5 frequencies are generated for each motifs: motif matches vs. all matches (by all motifs included in the analysis) (A), motif matches vs. matches of this particular motif (B), and target sequences (or accessions) with at least one motif match vs. all input sequences (C), vs. all matched sequences (by all motifs included in the analysis) (D) and vs. the sequences matched by the motif in question (E).

Next, each motif-driven information is classified regarding **single/multiple occurrence**, **plus/minus strands presence**, and **position** (when possible). And all these results are exhaustively detailed in a report file:

~/Desktop/Perl/cisAnalyzer/MyOutputFiles/cisA_user_date_Report.txt

The valuable feature of the report file resides in the possibility of browsing target sequences ids as they are grouped with other sequences sharing a particular category of motif presence.

Regarding localization information, the sequences are automatically divided in 2 or 4 zones when possible. This depends on the similarity of lengths within the target sequences set (forcing the consideration of the shortest sequence when they are different) and on the length comparison of the longest valid motif to be searched with the shortest target sequence. This consideration is schemed for each run in the report file, when browsing zone-specific information.

Co-occurrence analyses are performed when working with 2, 3 or 4 motifs due to particular output files limitations. All combinations of motifs are now considered and the information from each cis-element is intersected to obtain the results. The user could browse the co-occurrence results and target id data from the report file.

A particular default result that the user could quickly found is driven from the relative frequencies cisAnalyzer work with. It is the case of **enrichment** results. A particular file is generated to rapidly obtain those overrepresented motifs in the set of input target sequences.

~/Desktop/Perl/cisAnalyzer/MyOutputFiles/cisA_user_date_Enrichment Report.xls

The user could also go to the report file to search for particular enrichment frequencies.

Besides the mentioned reports, the most important feature of cisAnalyzer regarding outputs is the representation of frequencies results in 2 (A and C, for one motif analyses) or 5 (A to E, for more than one motif analyses) different graph files which handle one of both types of relativizations (and refer to

the above mentioned denominations of relativizations). They are generated by Perl system calls to R, executing particular /tmp located R scripts that use matrix generated by cisAnalyzer as input. The graphs consist of heatmaps that represent each category-related frequency for each searched motif, allowing a rapid visualization of motif-centered data that characterize the set of target sequences. Moreover, a hierarchical clustering of the represented motifs could be seen, pointing similarities between the found motifs regarding their presence footprint in the set of targets. Graph A (for one motif analyses) or A and B (for more than one motif analyses) show matches frequencies while C (for one motif analyses) or C, D and E (for more than one motif analyses) show frequencies of sequences with at least one match. For example, enrichment results are evidenced in graphs with relative frequencies of target sequences with at least one match of a given motif.

~/Desktop/Perl/cisAnalyzer/MyOutputFiles/cisA_user_date_Results Graph A.png

~/Desktop/Perl/cisAnalyzer/MyOutputFiles/cisA_user_date_Results Graph B.png

~/Desktop/Perl/cisAnalyzer/MyOutputFiles/cisA_user_date_Results Graph C.png

~/Desktop/Perl/cisAnalyzer/MyOutputFiles/cisA_user_date_Results Graph D.png

~/Desktop/Perl/cisAnalyzer/MyOutputFiles/cisA_user_date_Results Graph E.png

Another point of consideration in user's decisions but also in outputs generation is motif quantity as it compromises output files quality. One motif frequency results are represented with bar charts, 2 to 4 motifs frequencies are showed with default co-occurrence analysis in output heatmaps, 5 to 40 motifs' frequencies could be seen in proper heatmaps without co-occurrence analysis, and more than 40 motifs analysis must be cut for showing only the 40 cis-elements with the highest occurrence in the input set of targets.

Other extra output files are a heatmap of frequencies associated to groups of motifs which is obtained only when found PASPdb motif are included in more than one PASPdb group, and the legends of the generated graphs.

~/Desktop/Perl/cisAnalyzer/MyOutputFiles/cisA_user_date_Groups Graph.png

~/Desktop/Perl/cisAnalyzer/MyOutputFiles/cisA_user_date_Legends.txt

~/Desktop/Perl/cisAnalyzer/MyOutputFiles/cisA_user_date_Legends2.txt

Finally, Run_cisAnalyzer.pl execution ending is evidenced when it displays the occurrence or not of pattern matching(s) on terminal. Users should verify generated output files at /MyOutputFiles folder and interpret the biological significance of the obtained data.

cisAnalyzer initial tests

In order to verify proper cisAnalyzer working, we propose three different analyses and offer the output files they generate. These analyses are also proofs of cisAnalyzer capabilities including position, cooccurrence and motif discovery analyses.

1) Unknown Motif Discovery analysis

The file ~/Desktop/Perl/cisAnalyzer/Test files/UNK_test_targets.txt has all the needed features as a target sequences file and a fast observation of its sequences allows seeing a particular enriched motif (ACCGAC) in this set of *A. thaliana* and fictitious *R. phalange* species.

To obtain desired results, the user must run Start_cisAnalyzer.pl program, choose UNKNOWN MOTIF ENRICHMENT (U) - DNA (D) - Any DNA sequences (1) - Target sequences with the same length (1). Then, verify that /MyTargets folder is empty and insert UNK_test_targets.txt .txt file from /Test files folder. Later, the user must declare not working with known Phytozome species (as *R. phalange* is not) and then enter that there is 1 non-Phytozome species within the set of targets. After that, typing in its exact name Rphalange (as it is depicted in multifasta file) will allow the test to run properly. After Start_cisAnalyzer.pl running, the program informs the following:

The test proof went OK! YOU ARE READY TO USE cisAnalyzer!!

Checking /MyOutputFiles folder the user could see input and removed sequences, and the file of motif discovery (also included in /Test files folder as cisA_UNK_test_Motif discovery.txt) which must have been recognized the enrichment of ACCGAC and submotifs. These motifs are also searched in PASPdb and the results show they are different AP2/ERF TFs (including DREBs, RAP2, and TINY2 from several groups) binding sites.

Additionally, the analysis serves as an example of the type of target sequences that the user could input, what happens when inputting sequences with incorrect characters, and the requirements for inputting sequences from non-Phytozome species as *R. phalange*.

2) Cooccurrence analysis of custom aminoacidic motifs

At ~/Desktop/Perl/cisAnalyzer/Test files/ folder the user could find Protein_test_motifs.txt and Protein_test_targets.txt files, that allow performing a custom search of given motifs in the target sequences multifasta file and also a cooccurrence and position analysis between found motifs. Both files serve as examples of needed formats for proper cisAnalyzer functioning.

To obtain desired results, the user must run Start_cisAnalyzer.pl program, choose CUSTOM MOTIF ANALYSIS (C) - Protein (P) - Target sequences with the same (1) or different length (2). Then, verify that /MyTargets folder is empty and insert Protein_test_targets.txt .txt file from /Test files folder and Protein_test_motifs.txt .txt file at /MyMotifs emptied folder. The target sequences belong to a known plant species. Then, executing Run_cisAnalyzer.pl program will let the user to obtain desired results at /MyOutputFiles folder. Among the output files, Graph C shows enrichment frequencies of targets with at least one match of a given motif vs. input targets. This file is also included in /Test files folder as cisA_GraphC_Protein_test.png in order to compare and verify correct

cisAnalyzer functioning and to reveal the potentiality of these type of analyses, finding cooccurrence of more than one motif in particular positions of target sequences.

3) Cooccurrence analysis of custom ribonucleotidic motifs

In a similar way, using RNA_test_motifs.txt and RNA_test_targets.txt files from ~/Desktop/Perl/cisAnalyzer/Test files/ folder would help the user to verify the proper output generation of a custom ribonucleotidic motif analysis on RNA sequences. The user should proceed inputting motif and target sequences files in proper folders and choose CUSTOM MOTIF ANALYSIS (C) - RNA (R) during Start_cisAnalyzer.pl execution. A new category could be seen in output files as nucleotidic sequences allow strand classification of motifs occurrence. cisA_GraphC_RNA_test.png is also included in /Test files folder for comparisons with generated outputs.

References

- Artlip, T.S., Wisniewski, M.E., Bassett, C.L. and Norelli, J.L.** (2013) CBF gene expression in peach leaf and bark tissues is gated by a circadian clock. *Tree Physiol.*, **33**, 866–877.
- Brkljacic, J. and Grotewold, E.** (2017) Combinatorial control of plant gene expression. *Biochim. Biophys. Acta - Gene Regul. Mech.*, **1860**, 31–40. A
- Cheng, M.-C., Liao, P.-M., Kuo, W.-W. and Lin, T.-P.** (2013) The Arabidopsis ETHYLENE RESPONSE FACTOR1 regulates abiotic stress-responsive gene expression by binding to different cis-acting elements in response to different stress signals. *PLANT Physiol.*, **162**, 1566–1582.
- Chow, C.-N., Zheng, H.-Q., Wu, N.-Y., et al.** (2016) PlantPAN 2.0: an update of plant promoter analysis navigator for reconstructing transcriptional regulatory networks in plants. *Nucleic Acids Res.*, **44**, D1154–D1160.
- Farinati, S., Rasori, A., Varotto, S. and Bonghi, C.** (2017) Rosaceae Fruit Development, Ripening and Post-harvest: An Epigenetic Perspective. *Front. Plant Sci.*, **8**.
- Genero, M., Gismondi, M., Monti, L.L., Gabilondo, J., Budde, C.O., Andreo, C.S., Lara, M. V, Drincovich, M.F. and Bustamante, C.A.** (2016) Cell wall-related genes studies on peach cultivars with differential susceptibility to woolliness: looking for candidates as indicators of chilling tolerance. *Plant Cell Rep.*, **35**, 1235–1246.
- Pons, C., Martí, C., Forment, J., Crisosto, C.H., Dandekar, A.M. and Granell, A.** (2014) A Bulk Segregant Gene Expression Analysis of a Peach Population Reveals Components of the Underlying Mechanism of the Fruit Cold Response J.-S. Zhang, ed. *PLoS One*, **9**, e90706.
- Sebastian, A. and Contreras-Moreira, B.** (2014) footprintDB: a database of transcription factors with annotated cis elements and binding interfaces. *Bioinformatics*, **30**, 258–265.
- Sun, S., Yu, J.-P., Chen, F., Zhao, T.-J., Fang, X.-H., Li, Y.-Q. and Sui, S.-F.** (2008) TINY, a Dehydration-responsive Element (DRE)-binding Protein-like Transcription Factor Connecting the DRE- and Ethylene-responsive Element-mediated Signaling Pathways in Arabidopsis. *J. Biol. Chem.*, **283**, 6261–6271.
- Wang, K., Yin, X.-R., Zhang, B., Grierson, D., Xu, C.-J. and Chen, K.-S.** (2017) Transcriptomic and metabolic analyses provide new insights into chilling injury in peach fruit. *Plant. Cell Environ.*, **40**, 1531–1551.
- Wisniewski, M.E., Bassett, C.L., Renaut, J., Farrell, R., Tworkoski, T. and Artlip, T.S.** (2006) Differential regulation of two dehydrin genes from peach (*Prunus persica*) by photoperiod, low temperature and water deficit. *Tree Physiol.*, **26**, 575–584.

- Yilmaz, A., Mejia-Guerra, M.K., Kurz, K., Liang, X., Welch, L. and Grotewold, E.** (2011) AGRIS: the Arabidopsis Gene Regulatory Information Server, an update. *Nucleic Acids Res.*, **39**, D1118–D1122.
- Zhou, Y., Zhou, H., Lin-Wang, K., Vimolmangkang, S., Espley, R. V., Wang, L., Allan, A.C. and Han, Y.** (2014) Transcriptome analysis and transient transformation suggest an ancient duplicated MYB transcription factor as a candidate gene for leaf red coloration in peach. *BMC Plant Biol.*, **14**, 388.