Mele *BMC Bioinformatics* (2016) 17:50
DOI 10.1186/s12859-016-0896-x

BMC Bioinformatics

SOFTWARE Open Access

Arabidopsis Motif Scanner

Giovanni Mele

Abstract

Background: The major mechanism driving cellular differentiation and organism development is the regulation of gene expression. *Cis*-acting enhancers and silencers have key roles in controlling gene transcription. The genomic era allowed the transition from single gene analysis to the investigation of full transcriptomes. This transition increased the complexity of the analyses and the difficulty in the interpretation of the results. In this context, there is demand for new tools aimed at the creation of gene networks that can facilitate the interpretation of Next Generation Sequencing (NGS) data.

Results: Arabidopsis Motif Scanner (AMS) is a Windows application that runs on local computers. It was developed to build gene networks by identifying the positions of *cis*-regulatory elements in the model plant *Arabidopsis thaliana* and by providing an easy interface to assess and evaluate gene relationships. Its major innovative feature is to combine the *cis*-regulatory element positions, NGS and DNA Chip Arrays expression data, Arabidopsis annotations and gene interactions for the identification of gene networks regulated by transcription factors. In studies focused on transcription factors function, the software uses the expression data and binding site motifs in the regulative gene regions to predict direct target genes. Additionally, AMS utilizes DNA-protein and protein-protein interaction data to facilitate the identification of the metabolic pathways regulated by the transcription factor of interest.

Conclusions: Arabidopsis Motif Scanner is a new tool that helps researchers to unravel gene relations and functions. In fact, it facilitates studies focused on the effects and the impact that transcription factors have on the transcriptome by correlating the position of cis-acting elements, gene expression data and interactions.

Keywords: Arabidopsis, *cis*-elements, Binding motif, Gene network, Next Generation Sequencing, DNA Chip Array, Gene expression

Background

The recent advent of the genomic era, first with the DNA Chip Arrays then with Next Generation Sequencing (NGS) technologies, revolutionized the classical view that divided the genome into two entities: one comprising genes and their regulatory regions responsible for encoding messenger RNA translated in turn into protein; and one comprising “junk DNA” with unknown and consequently nonessential function [1]. By NGS, it was possible to identify that non-coding transcripts (ncRNA), which originate from intergenic sequences previously defined as junk, far exceed those of protein-coding genes [2]. This led to the discovery of novel layers of complexity in gene organization and expression highlighting the extreme versatility of genomes [3, 4]. Moreover, the genomic era has radically modified the entire field of biology by changing the approach and the challenges that biologists have to face [5–8]. Indeed, the advent of NGS technologies has enabled researchers to move from single gene analysis to the investigation of full transcriptomes. Although this transition allows one to have a global view on the gene expression changes, it came at a price. On one side, the enormous amount of data increased the complexity of the analyses and of the management of information; on the other side, it increased the difficulties in the interpretation of the results obtained. In the near future the major challenge in software development will be platforms that can best integrate the NGS results to facilitate data interpretation. This new generation of programs will be considered successful when they will provide an extensive view on gene

Correspondence: melegio@ibba.cnr.it
Institute of Agricultural Biology and Biotechnology (IBBA), National Council of Research (CNR), Via Salaria Km. 29.300, Monterotondo Scalo (Roma) 00015, Italy

interactions and molecular pathways that can help the comprehension of gene relations and functions.

Cellular differentiation and organism development are governed by precise gene expression patterns. The diverse expression patterns in the different cells are established by the coordinated action of intergenic, as well as intragenic, *cis*-acting enhancers and silencers known as *cis*-regulatory elements [9–11]. *Cis*-acting elements such as core and proximal promoter elements are typically restricted to within a couple of hundred base pairs from transcriptional start sites and regulate genes in their immediate vicinity. In contrast, distal *cis*-elements are usually located at >1 kb and in some cases up to 1 Mb in either direction from a transcription start site. Functional DNA sequences change at a lower rate over evolutionary time than sequences without function [12, 13]. Consequently, *cis*-regulatory elements tend to be conserved, whereas functionless sequences are randomized by substitution, lost by conversion, or deleted entirely. In this context, genome-wide studies of *cis*-regulatory elements become the key path to build metabolic pathways and gene networks for a better comprehension of gene relations and transcription factors function. In fact, the first step to unravel the function of a transcription factor is the identification of the *cis*-regulatory element that it binds and the target genes under its control. Subsequently, the clusterization of the target genes in networks allows the identification of the metabolic pathways regulated by the transcription factor of interest and consequently of its function.

To date, several different platforms for expression data analysis and management have been developed; however, there is a demand for software for Arabidopsis data interpretation. Most of the available promoter analysis software focuses on the presence of well-characterized *cis*-acting elements in a single user provided promoter region. Alternatively, they scan the Arabidopsis genome and provide a list of loci where the *cis*-acting element is present. AMS facilitates transcription factors function identification by implementing data interpretation. In fact, AMS allows the search of *cis*-acting elements, the organization of the differentially expressed target gene data and the identification of gene networks.

Implementation

Arabidopsis Motif Scanner (AMS) executable is freely available on the web page of the Institute of Agricultural Biology and Biotechnology of National Council of Research (http://www.ibba.mlib.cnr.it/Arabidopsis_Motif_Scanner.html) and at SourceForge open-source repository (http://sourceforge.net/projects/arabidopsismotifscanner/files/?source=navbar). This software was developed in C# language and was designed to be fully compatible with Windows 7, 8 and 10 environments. Arabidopsis Motif Scanner is a user-friendly application developed for the Windows environment to be fully compatible with the Illumina NGS and the Affymetrix Gene Chip platforms. Arabidopsis Motif Scanner GUI consists of one window with four tabs: *Motif Scanner*, *Expression*, *Gene Viewer* and *Interactions* (Fig. 1).

Motif scanner Tab

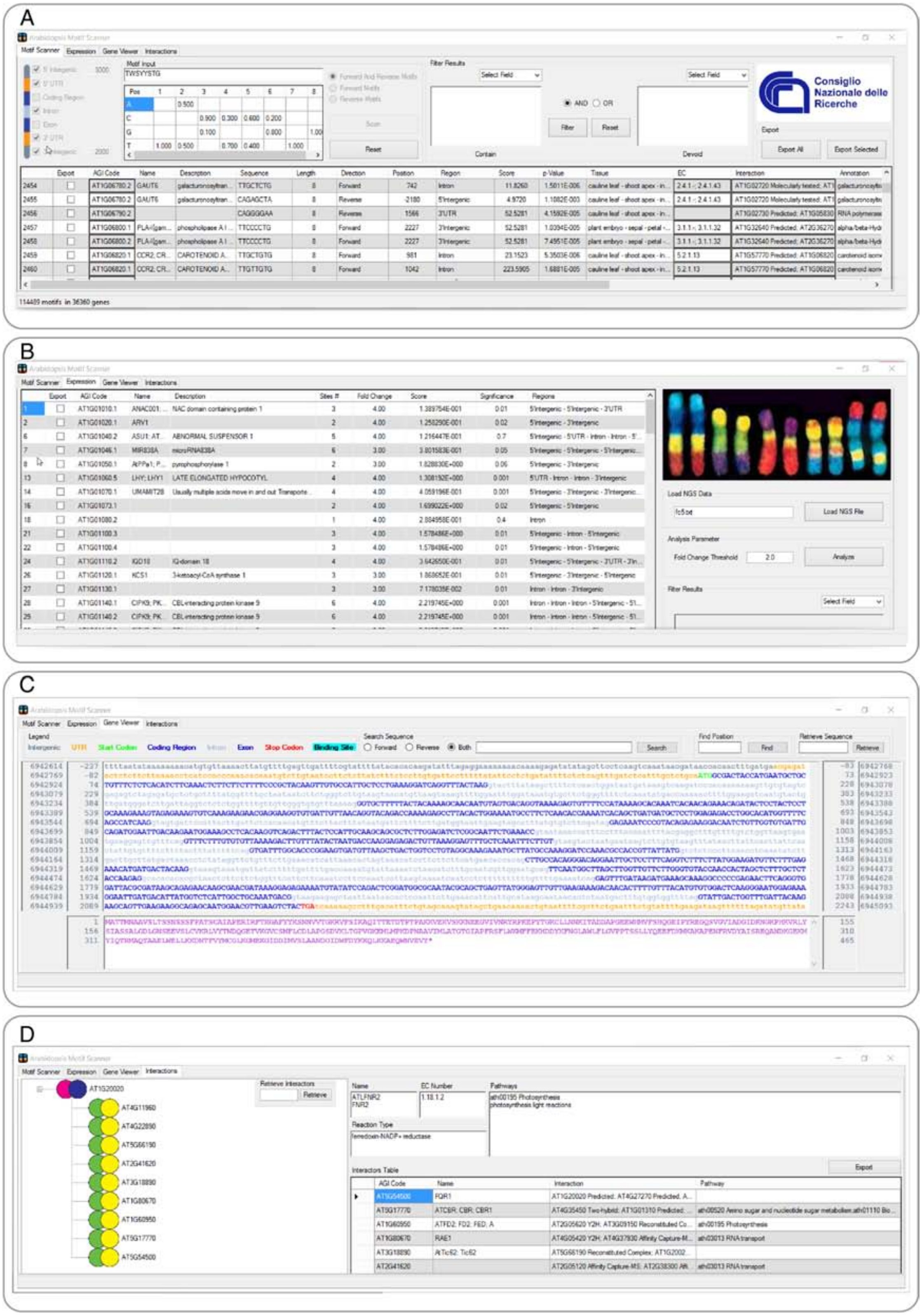
The Arabidopsis Motif Scanner starts up by loading the database and displays the *Motif Scanner* tab (Fig. 1a). The tab consists of an upper setting mask dedicated to the input of working parameters and a lower mask that shows the results.

The setting mask includes four panels (from left to right): *chromosome region selection*; *motif input*; *filter function* and *data export*. In the *chromosome region selection* panel, the tick boxes allow the selection of chromosome regions that will be included in the analysis. The rectangular boxes, which flank the 3' and 5' intergenic tick boxes, provide analysis flexibility by allowing the user to select the intergenic sequence length. If the intergenic length boxes are left blank, the software considers as intergenic the regions between genes. In the case that intergenic sequences are shorter than the user input, the intergenic regions will be considered up to the flanking coding sequence.

The *motif input* panel contains the user input sequence box and allows the search for the motif in forward, reverse or both orientations. The software accepts standard nucleotide degeneration (Table 1) and an editable Position-specific Weight Matrix (PWM) allows a refined and flexible search. The PWM expresses the probability that the transcription factor of interest binds the motif considered. For transcription factors with a known binding motif sequence the PWM matrix can be easily retrieved from the literature (PWM sequence logo), while for transcription factors with unknown binding motif sequence the PWM matrix is obtained experimentally by the Selection And Amplification Binding (SAAB) assay.

To obtain a reasonable number of output hits for the subsequent data interpretation, a motif of at least 7 bp long is suggested. To refine the outputs, a filter can be set; the selective parameters are typed in the *Contain* and *Devoid* windows to function in parallel or in exclusive way (using the “and/or” radio buttons). The *data export* panel consents the export of results in a tab delimited text file. The tickable cells of the export column in the results table allows a selective export of the data of interest.

The lower mask reports a column-sortable table which visualizes AGI Code, Name of the gene, Description of gene, Sequence of the targeted motif, Length of the



Motif Input: TWSYYSTG

Pos	1	2	3	4	5	6	7	8
A		0.500						
C			0.900	0.300	0.600	0.200		
G			0.100			0.800		1.00
T	1.000	0.500		0.700	0.400		1.000	

	Export	AGI Code	Name	Description	Sequence	Length	Direction	Position	Region	Score	p-Value	Tissue	EC	Interaction	Annotation
2454	□	AT1G06780.2	GAUT6	galacturonosyltran...	TTGCTCTG	8	Forward	742	Intron	11.8260	1.5011E-006	cauline leaf - shoot apex - in...	2.4.1.-; 2.4.1.43	AT1G02720 Molecularly tested; AT1	galacturonosyltr
2455	□	AT1G06780.2	GAUT6	galacturonosyltran...	CAGAGCTA	8	Reverse	-2180	5'Intergenic	4.9720	1.1082E-003	cauline leaf - shoot apex - in...	2.4.1.-; 2.4.1.43	AT1G02720 Molecularly tested; AT1	galacturonosyltr
2456	□	AT1G06790.2			CAGGGGAA	8	Reverse	1566	3'UTR	52.5281	4.1592E-005	cauline leaf - shoot apex - in...		AT1G02730 Predicted; AT1G05830	RNA polymerase
2457	□	AT1G06800.1	PLA-I{gam...	phospholipase A I ...	TTCCCCTG	8	Forward	2227	3'Intergenic	52.5281	1.0394E-005	plant embryo - sepal - petal -...	3.1.1.-; 3.1.1.32	AT1G32640 Predicted; AT2G36270	alpha/beta-Hydr
2458	□	AT1G06800.2	PLA-I{gam...	phospholipase A I ...	TTCCCCTG	8	Forward	2227	3'Intergenic	52.5281	7.4951E-005	plant embryo - sepal - petal -...	3.1.1.-; 3.1.1.32	AT1G32640 Predicted; AT2G36270	alpha/beta-Hydr
2459	□	AT1G06820.1	CCR2; CR...	CAROTENOID A...	TTGCTGTG	8	Forward	981	Intron	23.1523	5.3503E-006	cauline leaf - shoot apex - in...	5.2.1.13	AT1G57770 Predicted; AT1G06820	carotenoid isome
2460	□	AT1G06820.1	CCR2; CR...	CAROTENOID A...	TTGTTGTG	8	Forward	1042	Intron	223.5905	1.6881E-005	cauline leaf - shoot apex - in...	5.2.1.13	AT1G57770 Predicted; AT1G06820	carotenoid isome

114409 motifs in 36360 genes

	Export	AGI Code	Name	Description	Sites #	Fold Change	Score	Significance	Regions
1	□	AT1G01010.1	ANAC001; ...	NAC domain containing protein 1	3	4.00	1.389754E-001	0.01	5'Intergenic - 5'Intergenic - 3'UTR
2	□	AT1G01020.1	ARV1		2	4.00	1.258290E-001	0.02	5'Intergenic - 3'Intergenic
6	□	AT1G01040.2	ASU1; AT...	ABNORMAL SUSPENSOR 1	5	4.00	1.216447E-001	0.7	5'Intergenic - 5'UTR - Intron - Intron - 5'...
7	□	AT1G01046.1	MIR838A	microRNA838A	6	3.00	3.801583E-001	0.05	5'Intergenic - 5'Intergenic - 5'Intergenic...
8	□	AT1G01050.1	AtPPa1; P...	pyrophosphorylase 1	2	3.00	1.828830E+000	0.06	5'Intergenic - 3'Intergenic
13	□	AT1G01060.5	LHY; LHY1	LATE ELONGATED HYPOCOTYL	4	4.00	1.308192E+000	0.001	5'UTR - Intron - Intron - 3'Intergenic
14	□	AT1G01070.1	UMAMIT28	Usually multiple acids move in and out Transporte...	4	4.00	4.059196E-001	0.001	5'Intergenic - 3'Intergenic - 3'Intergenic...
16	□	AT1G01073.1			2	4.00	1.699022E+000	0.02	5'Intergenic - 5'Intergenic
18	□	AT1G01080.2			1	4.00	2.884958E-001	0.4	Intron
21	□	AT1G01100.3			3	4.00	1.578486E+000	0.01	5'Intergenic - Intron - 5'Intergenic
22	□	AT1G01100.4			3	4.00	1.578486E+000	0.01	5'Intergenic - Intron - 5'Intergenic
24	□	AT1G01110.2	IQD18	IQ-domain 18	4	4.00	3.642650E-001	0.01	5'Intergenic - 5'Intergenic - 3'UTR - 3'In...
26	□	AT1G01120.1	KCS1	3-ketoacyl-CoA synthase 1	3	3.00	1.868652E-001	0.01	5'Intergenic - 3'Intergenic - 5'Intergenic
27	□	AT1G01130.1			3	3.00	7.178035E-002	0.01	Intron - Intron - 3'Intergenic
28	□	AT1G01140.1	CIPK9; PK...	CBL-interacting protein kinase 9	6	4.00	2.219745E+000	0.001	Intron - Intron - Intron - 5'Intergenic - 5'I...
29	□	AT1G01140.2	CIPK9; PK...	CBL-interacting protein kinase 9	6	4.00	2.219745E+000	0.001	Intron - Intron - Intron - 5'Intergenic - 5'I...

Load NGS Data: fc5.txt

Fold Change Threshold: 2.0

Name: ATLFNR2
FNR2

EC Number: 1.18.1.2

Pathways: ath00195 Photosynthesis
photosynthesis light reactions

Reaction Type: ferredoxin-NADP+ reductase

Interactors Table

	AGI Code	Name	Interaction	Pathway
▶	AT5G54500	FQR1	AT1G20020 Predicted; AT4G27270 Predicted; A...	
	AT5G17770	ATCBR; CBR; CBR1	AT4G35450 Two-hybrid; AT1G01310 Predicted; ...	ath00520 Amino sugar and nucleotide sugar metabolism;ath01110 Bio...
	AT1G60950	ATFD2; FD2; FED; A	AT2G05620 Y2H; AT3G09150 Reconstituted Co...	ath00195 Photosynthesis
	AT1G80670	RAE1	AT4G05420 Y2H; AT4G37930 Affinity Capture-M...	ath03013 RNA transport
	AT3G18890	AtTic62; Tic62	AT5G66190 Reconstituted Complex; AT1G2002...	
	AT2G41620		AT2G05120 Affinity Capture-MS; AT2G38300 Affi...	ath03013 RNA transport

Fig. 1 Arabidopsis Motif Scanner Tabs. **a** *Motif Scanner* Tab. **b** Expression Tab. **c** *Gene Viewer* Tab. **d** *Interactions* Tab

Table 1 Degeneration Table. The first and third column report the character assigned to the degeneration. In the second and forth columns show the correspondent nucleotides

Deg	Nuc	Deg	Nuc
R	A, G	B	C, G, T
Y	C, T	D	A, G, T
S	G, C	H	A, C, T
W	A, T	V	A, C, G
K	G, T	N	A, G, C, T
M	A, C		

targeted motif, Region within the gene that holds the motif, Direction of the motif, Position with respect to the start codon, Motif Score, *p*-Value, Significance, EC Number (unique number assigned to a enzymatic reaction), Interaction of the gene product and the Annotation. Moreover, the AGI Code, EC and Interaction cells in the results table are clickable and respectively display gene organization inclusive of the binding motif positions, pathways in which the gene is involved and its protein interactions. The table further reports all matching motifs for all the different splicing forms. Finally, the total number of motifs appears at the bottom left corner.

Each Motif Score (*MS*) is calculated based on *observed:expected* frequency ratios (O/E) and takes into account the *PWM* specific motif value. Specifically, each *MS* is computed as:

$$MS = V_{PWM} * B_O * \ln\left({}^{B_O}/_{B_E}\right)$$

Where V_{PWM} expresses the probability that the transcription factor of interest binds that exact motif sequence considered and it is calculated on the PWM matrix created by the user. B_O and B_E represent the number of Binding sites Observed and Expected for the motif of interest respectively. B_E is calculated as:

$$B_E = M_P * T$$

M_p is the probability to observe the motif and depends on the motif sequence composition and the nucleotide frequencies in the genome (for Arabidopsis A,T = 0.32 and C,G = 0.18). M_p is calculated as the frequency of each nucleotide that constitutes the motif (e.g. M_P for ATCCG = 0.32x0.32x0.18x0.18x0.18). T represents the total number of possible matching positions for the input motif across both strands of the genome and it is calculated as:

$$T = 2 * (R_L - M_L + 1)$$

Where *2* accounts for two strands. R_L is the length of the DNA region where the motif occurs. M_L is the motif length.

The *p*-Value for each motif represents the probability of obtaining at least n motifs in the sequence and it is calculated by the cumulative binomial distribution.

$$\text{p–Value} = 1 - \sum_{i=0}^{n-1} \binom{T}{i} * (M_P)^i * (1 - M_P)^{T-i}$$

Where n is the number of motifs in the sequence. M_p is the probability to observe the motif and T represents the total number of possible matching positions.

Expression Tab

The Expression tab (Fig. 1b) allows the integration between *cis*-acting element positions and NGS or DNA Chip Array expression data. In the upper right section of the panel, a tab-delimited file containing the AGI code and the correspondent fold Change can be loaded. The Fold Change Threshold (FCT) box allows the user to set the best FCT for the analysis. The lower right box permits filtering the result. In the left section, a column-sortable table visualizes the AGI code Name, Description, Number of Sites, Fold Change, Regions, Binding Score and Significance for each gene.

The Binding Score (BS) is computed as:

$$BS = \sum_{i=1}^{n} V_{PWM} * B_{FC\ spec} * \ln\left(\frac{{}^{B_{FC\ Spec}}/_{B_{Tot\ Spec}}}{{}^{B_{FC}}/_{B_{Tot}}}\right)$$

Where n symbolizes the number of binding motifs present in each user-defined locus (genomic regions selected for the analysis). V_{PWM} expresses the probability that the transcription factor of interest binds that exact motif sequence considered and it depends on the PWM matrix created by the user. B_{FC} and $B_{FC\ Spec}$ represent the occurrence of binding sites above the fold change threshold for all the motifs derived from the PWM and for one specific motif, respectively. The fold change threshold is set by the user in the FCT box in the "Expression Tab" and depends only on the expression data quality. Finally, the B_{tot} and $B_{tot\ Spec}$ indicate the total motif number occurring in the genome for all the motifs considered derived from the PWM and for one specific motif, respectively. Moreover, the chi-squared test is used to determine whether there is a significant difference between the number of genes that contain binding sites and the number of genes that both are differentially expressed and contain binding sites for all the motifs derived from the PWM and for the one specific motif considered.

In the studies focused on the function of a transcription factor, the comparison between wild-type sample and mutant that over or under express the transcription factor of interest is a key step for the identification of the target genes. In this type of studies it is informative to distinguish between target genes which are directly bound by the transcription factor that alters their expression (primary targets) from genes with a mutated expression due to secondary effects (secondary targets). In AMS, the possibility to combine binding site positions with changes of expression (expressed as fold change) allows the discrimination between primary and secondary target genes.

Gene viewer Tab

The *Gene Viewer* tab (Fig. 1c) shows the gene organization by either clicking on the AGI code box of the *Motif Scanner* tab or by typing the AGI code in the *Retrieve Sequence Box* (top right). The upper and bottom sections include the gene structure and the protein sequence, respectively. The different gene regions appear colored according to the legend (upper left corner). Both left and right external columns report the absolute nucleotide position in the chromosome, while the internal columns include the positions with respect to the ATG codon start. The user can search for specific nucleotide sequences or positions by typing the *Search Sequence* and *Fine Position* functions. In the bottom section, the protein translation is displayed and is flanked by two columns, which report the aminoacid position with respect to the initial Methionine. Finally, the length and position of a selected sequence appear in the lower left corner.

Interactions Tab

The *Interactions* tab is displayed (Fig. 1d) after clicking on either EC or Interaction cells of the *Motif Scanner* tab. A tree view of the interactions appears in a window on the left side. Upon selection, the yellow and green interaction symbol turns blue and fuchsia showing the biochemically tested and computational assigned interactors of the gene of interest. Concerning the biochemically tested interactions, the AMS database contains information of protein-protein interaction, Protein Complementation Assay, Far Western Blotting, Pull Down, biochemical activity, Affinity Capture-Western, Yeast Two-Hybrid, Pull-Down Assay, Affinity Capture-MS, Co-Immunoprecipitation,in vitro Binding Assay, Split-Reporter Assay and phenotypic suppression and enhancement.

The right side of the *Interactions* tab consists of two panels, the top one reports the Name, EC number, Reaction Type and Pathways referred to the targeted gene, while, the bottom panel provides more information on the interactors, including the AGI Code, Name, Interactions and Pathway of the interactors. The tree view and the interactors table are synchronized for an easy consultation.

Database

The database behind the Arabidopsis Motif Scanner was built using several different sources. Arabidopsis genome organization and annotations was derived from "The Arabidopsis Information Resource" (TAIR) consortium (http://www.arabidopsis.org) and elaborated to best fit the requirements of the software. Biologically tested and computational derived interactions were retrieved from: 1) AtPID database (Jian C., *et al*. 2007) (http://www.megabionet.org/atpid/webfile); 2) the database published by Jane G. L. (Jane G. L., *et al* (2007); 3) the Plant Interactome Database resulted from a collaboration between the Salk Institute (http://signal.salk.edu/interactome/index2.html) and the Center for Cancer Systems Biology (http://interactome.dfci.harvard.edu/A_thaliana/index.php); 4) BioGRID (http://www.thebiogrid.org/); 5) IntAct of EMBL-EBI (http://www.ebi.ac.uk/intact/); 6) BIND (http://bind.ca/); 7) TAIR.

Results and discussion

The majority of the web interfaces available on the net are developed to perform analyses of *cis*-acting elements on human and animal genomes. Moreover, several of them are commercial platforms and analyses are typically expensive. As for the plant kingdom, the web interfaces were developed for Arabidopsis since, as model plant, it has the best-annotated genome. Web interfaces such as PLACE (https://sogo.dna.affrc.go.jp/cgi-bin/sogo.cgi?lang=en&pj=640&action=page&page=newplace) were developed based on non-holistic criteria, in other words, the analyses can be performed only for well-known *cis*-acting elements and are restricted to single promoters. On the contrary, the Arabidopsis Motif Scanner was developed to identify the positions of new and unknown *cis*-acting elements in the entire genome by using an easy-to-use GUI. The AMS software shows a few similarities with genome-scale-DNA-pattern match of Regulatory Sequence Analysis Tools (RSAT) (http://floresta.eead.csic.es/rsat/) and PatMatch of The Arabidopsis Information Resource (TAIR) (http://www.arabidopsis.org/cgi-bin/patmatch/nph-patmatch.pl) web-based GUIs. Differently from those, AMS provides gene annotation and information on gene interactions, which are essential for the identification and construction of gene networks and integration with NGS expression data for the comprehension of new gene relations and functions.

Arabidopsis Motif Scanner is more flexible than PatMatch and RSAT in the choices of the genomic regions. For instance, Arabidopsis Motif Scanner allows choosing any combination of genomic regions while PatMatch has some limitation not allowing analyzing upstream and downstream regions in the same run. Concerning RSAT, it does not permit searches in multiple regions. The AMS flexibility allows identifying motifs falling across two contiguous regions that it is not possible to be identified with both PatMatch and RSAT. Moreover, PatMatch poses limits on the intergenic regions length (500 or 1000 or 3000 bp) and does not take into account if the intergenic regions are shorter because of the presence of coding sequences of flanking genes. These limitations are overcome in AMS that consider intergenic region the portion of sequence between two coding sequences. Finally, AMS offers the option to use PWM for the input motif differently from PatMatch and RSAT.

Most genes encode multiple transcripts by alternative promoters, alternative splicing, or alternative polyadenylation [14–18]. The combinatorial mechanism of alternative splicing increases the coding potential of the genome by allowing the synthesis of multiple protein isoforms with different—even antagonistic—functions from a single gene [19]. AMS offers the advantage to analyze the motif occurrence in multivariate splicing forms, providing information on why differentially spliced transcripts are found in the diverse tissues and organs. Moreover, the Gene Viewer Tab is an easy tool to visualize motif positions and organization for the genes of interest. Contrarily to PatMatch and RSAT, AMS reports the motif in the 5' regions as "negative" positions with respect to the ATG start. One peculiar feature of AMS is the possibility to combine *cis*-acting element search and Next Generation Sequencing data for a better comprehension of gene relations and functions.

AMS becomes particularly useful for the studies of transcription factor (TF) function. To this aim, two aspects must be considered to an effective result. First, it is mandatory to identify the primary targeted genes of the TF and their respective functions. Second, it is necessary to detect transcriptome differences between wild-type plants versus mutant, which is affected in a given TF. AMS output consists of a gene list that match motif occurrence and intensity of differential expression allowing a significant selection of putative primary targets. The list of this primary targets combined with gene annotations and interaction further provides the information on molecular pathways in which a TF exerts functions.

Conclusion

Arabidopsis Motif Scanner is a powerful user-friendly tool that runs on local computers allowing correlation of genomic–wide searches for *cis*-acting element position and NGS or DNA Chip Array expression data. The great advantage of the software, that distinguishes it from other web interfaces, is the presence of a database that provides annotations and gene interactions for the hits. When combined with expression data, this software enhances the interpretation of NGS and DNA Chip Array results and allows the discovery of new gene relations and functions. Moreover, Arabidopsis Motif Scanner can be considered an effective tool to identify new functions of transcription factors. In fact, the genome-wide screen of transcription factors binding motifs provides valuable information on the probable target genes and consequently on the metabolic pathways in which the transcription factor of interest is involved.

Availability and requirements

Project name: Arabidopsis Motif Scanner

Project home page: http://www.ibba.mlib.cnr.it/Arabidopsis_Motif_Scanner.html

Operating system(s): Windows

Programming language: C#

Any restrictions to use by non-academics: license needed: GNU General Public License version 3.0 (GPLv3)

Competing interests

The author declares that he has no competing interests.

Acknowledgements

Special thanks to Giulo Testone for helping with the statistic and to Donato Giannino for suggestions, comments and the critical reading of the paper. Thanks to Patrick O'Keeffe for the English critical review. This work was supported by a dedicated grant from the Italian Ministry of Economy and Finance to the National Research Council for the project "Innovazione e Sviluppo del Mezzogiorno - Conoscenze Integrate per Sostenibilità ed Innovazione del Made in Italy Agroalimentare - Legge n. 191/2009"

Received: 11 February 2015 Accepted: 19 January 2016
[redacted]

References

1. Susumo O. So much "junk" DNA in our genome. Smith HH, editor Evolution of genetic systems New York: Gordon and Breach 1972:366–370.
2. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74.
3. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, et al. Architecture of the human regulatory network derived from ENCODE data. Nature. 2012;489(7414):91–100.
4. Pennisi E. Genomics. ENCODE project writes eulogy for junk DNA. Science. 2012;337(6099):1159–61.
5. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009;10(10):669–80.
6. Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. Nat Rev Genet. 2010;11(7):476–86.
7. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010;11(1):31–46.
8. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. Nat Rev Genet. 2011;12(2):87–98.
9. Mejia-Guerra MK, Pomeranz M, Morohashi K, Grotewold E. From plant gene regulatory grids to network dynamics. Biochim Biophys Acta. 2012;1819(5): 454–65.
10. Plank JL, Dean A. Enhancer function: mechanistic and genome-wide insights come together. Mol Cell. 2014;55(1):5–14.
11. Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet. 2014;15(4):272–86.
12. Freeling M, Rapaka L, Lyons E, Pedersen B, Thomas BC. G-boxes, bigfoot genes, and environmental response: characterization of intragenomic conserved noncoding sequences in Arabidopsis. Plant Cell. 2007;19(5): 1441–57.
13. Inada DC, Bashir A, Lee C, Thomas BC, Ko C, Goff SA, et al. Conserved noncoding sequences in the grasses. Genome Res. 2003;13(9):2030–41.

14. Missihoun TD, Kirch HH, Bartels D. T-DNA insertion mutants reveal complex expression patterns of the aldehyde dehydrogenase 3H1 locus in Arabidopsis thaliana. J Exp Bot. 2012;63(10):3887–98.
15. Shu H, Wildhaber T, Siretskiy A, Gruissem W, Hennig L. Distinct modes of DNA accessibility in plant chromatin. Nat Commun. 2012;3:1281.
16. Xing D, Li QQ. Alternative polyadenylation and gene expression regulation in plants. Wiley Interdiscip Rev RNA. 2011;2(3):445–58.
17. Chen WH, Lv G, Lv C, Zeng C, Hu S. Systematic analysis of alternative first exons in plant genomes. BMC Plant Biol. 2007;7:55.
18. Wu X, Liu M, Downie B, Liang C, Ji G, Li QQ, et al. Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation. Proc Natl Acad Sci U S A. 2011;108(30):12533–8.
19. Pose D, Verhage L, Ott F, Yant L, Mathieu J, Angenent GC, et al. Temperature-dependent regulation of flowering by antagonistic FLM variants. Nature. 2013;503(7476):414–7.