Pak. J. Bot., 51(4): 1297-1302, 2019. DOI: http://dx.doi.org/10.30848/PJB2019-4(44)

IDENTIFICATION AND ANALYSIS OF DNA BINDING SPECIFIC TRANSCRIPTION FACTOR` BINDING SITES IN SUCROSE SYNTHASE PROMOTER

HIRA MUBEEN[1], AMMARA MASOOD[1], JAVAID IQBAL WATTOO[1], AMMARA NASIM[2] AND SHAHID RAZA[3]

[1]University of Central Punjab, Lahore, Pakistan
[2]Government College University, Faisalabad, Pakistan
[3]Lahore Garrison University, Lahore, Pakistan
**Corresponding authors email: hira_sh@hotmail.com*

Abstract

Sucrose synthase (SUS) is one of the abundantly expressed genes in plants. In this study, the promoter sequence of SUS gene was identified, analyzed and retrieved from high throughput genomic sequence (HTGS) database. The transcription factor binding sites (TFBs) of subject promoter were identified through different bioinformatics tools. The analysis revealed several crucial motifs and TFBs within the entire promoter region. The most common site found within promoter region was AGL3, which binded with DNA-region specific binding site and involved in process of transcriptional regulation. AGL3 is also a key factor for studying protein dimerization activity. Furthermore, AGL3 encodes a protein, which encodes a conserved domain, MADS protein domain. The analysis of AGL3 through (Plant PAN 2.0) showed three associated TF: AT5G23260, AT2G03710, U81369, which were used for further analysis of conserved protein domains through Pfam and InterPro. We found that these TF could help in developmental regulation and validation of candidate genes. Moreover, the binding of AGL3 with serum response factor (SRF), a nuclear protein also indicates the transcriptional regulations of several growth factors. This core domain belongs to MADS protein family, which interacts with certain accessory factors and varies up to 90 amino acids. Analysis of such promoter sequences and their predicted TF can be useful to understand the transcriptional regulatory networks in plant genetic engineering.

Key words: AGL3, TF, Growth factors, SRF, MADS, SRF.

Introduction

Sucrose synthase (SUS) is an important enzyme essentially required for carbohydrate metabolism in plants. Sucrose synthase is actively involved in various metabolic, structural and storage functions. It can convert sucrose into fructose. In Arabidopsis, nearly six isoforms of sucrose synthase are found abundantly. All SUS genes carry different expression patterns (Baud *et al*., 2004; Fallahi *et al*., 2008). Sucrose synthase is uniquely able to mobilize sucrose into multiple pathways involved in metabolic, structural, and storage functions. Recent research findings indicates the biological function of SUS, which may extend beyond its catalytic activity. This inference is based on the following observations: (a) tissue-specific, isoform-dependent and metabolically regulated association of SUS with mitochondria and (b) isoform-specific and anoxia-responsive interaction of SUS with the voltage-dependent anion channel (VDAC), the major outer mitochondrial membrane protein. Recent research revealed the localization of both VDAC and SUS to the nucleus in maize seedling tissues. Their intricate regulation under anoxia indicates the importance of these two proteins in inter-compartmental signaling. However, analysis of SUS gene promoter for finding putative transcription factors and transcription factor binding sites has made it possible to study the structure and function. The analysis of SUS by using Plant PAN showed presence of various TFB's within the promoter region.

Proteins play an important role in gene expression and regulation. Gene expression is an essential feature to study the function and importance of particular proteins. Several proteins are involved in mechanism, which control gene expression and its regulatory processes. Some specialized proteins like transcription factors (TF's) are involved in control of some other genes which are turned on or off in the genome. Binding of a TF to its specific DNA binding site is a most important step to initiate and regulate the transcription of its target genes. These factors have specialized DNA binding domains, which are useful to transcribe DNA into RNA. They can also bind to specific DNA sequences known as promoters and interconnected by several regulatory networks. The majority of cellular process are dependent on expression of genes. The regulatory mechanism of genes results due to interactions among specialized proteins, namely, transcription factors. Several genes are connected in merged networks of thousands of transcription factors. However, the transcription factors consist of small transcription factor binding sites. The transcriptional regulation is a complex process mainly controlled by DNA-binding transcription factors (TFs) that enhance the gene expression by facilitating the binding of RNA polymerase with specific promoters.

All TFs have several different families depending specifically on two domains, each having different function as a regulatory switch. Out of these two domains, one functions in protein-protein interactions and other is involved in interacting and binding with target DNA sequences **(**Ptashne & Gann, 2002; Martinez-Antonio *et al*., 2002; Babu & Teichmann, 2003; Seshasayee *et al*., 2006). However, the use of modern experimental techniques helps to determine target genes and their related sites, but this is costly and time consuming (Stormo, 2000). However, algorithms can better describe transcription factor binding sites in a hierarchal manner (Yada *et al*., 1999). Moreover, the concept of the consensus sequence represents the specific binding of TF with TFBs. Similarly, the binding

sites for sucrose synthase promoter AGL3 are "GTTTGGTAGTATGGAGGA". Some of the potential binding sites can be determined by comparing consensus sequence with specific target sequence.

AGL3, the most abundant TFB was identified on the basis of sequence similarity to the floral homeotic gene AGAMOUS (AG).This gene encodes a specialized protein with a conserved domain, MADS domain. AGL3 is a well-known MADS-box gene with a general intron-exon structure, similar to many other plants MADS-box genes. The MADS-box genes has been identified in various eukaryotic genomes including, plants, insects and mammals **(**Messenguy & Dubois, 2003**)**. These are also represented as DNA-binding transcription factors. MADS box motif plays an important role as a member of MADS-box family of transcription factors. The DNA binding domain (DBD) interacts with specific DNA sequences. The DNA binding domain is mainly involved in bringing up the transcription activation domain near initiation complex. Most of these domains have three main functions. Initially, a) they bind with the major groove of DNA, b) they can affixed into minor groove and c) they can bring conformational changes in the double helix. The presence of TFs and TFBs is specific to a particular promoter sequence.

Promoters are regulatory regions of genes mainly involved in controlling overall expression profile of a gene. Plant promoters are mostly important for gene expression studies and for production of improved crops (Potenza *et al*., 2004). Transcription factors can easily bind to promoter near transcription start site that leads to formation of transcription initiation complex. Similarly, some other transcription factors bind to other regulatory regions such as enhancer regions. Transcriptional gene regulation is the most important feature to control important genes. The action of transcription factors permits a unique expression of each gene. TF's recognize some specialized transcription factor binding sites (TFB's) within the promoter region of gene. However, several conformational changes were observed in structure of transcriptional factors, dependent on binding patterns of DNA. However, the correct binding of transcription factors is essential for gene assembly and regulatory process. Furthermore, some of them function in a variable and different fashion by sending signals within the cell.

Plant promoters are useful for various agricultural research applications. The increased demand of promoters for improvement of transgenic plants is in practice. The regulation of plant genes is truly dependent on promoters. The expression of transgenic plants can be upregulated and down regulated by use of specific promoters. The promoters are categorized into three main types including, constitutive, inducible and tissue specific. Nowadays, a new type of synthetic promoters are being used to fulfil the needs of current research methods in genetic engineering. The synthetic promoters are more valuable to attain the targeted gene expression like D-hordein wheat promoter (Masood *et al.,* 2017). They are designed especially according to the need of subject gene and presence of conserved regions or motifs near upstream region of promoter. Newly designed promoters with specific characteristics, have gained much more importance in plant molecular biology research.

In the present study, we have identified transcription factor binding sites in Arabidopsis Sucrose synthase gene promoter. The analysis showed various DNA binding specific sites. We further characterized the protein domains present within the MADS-box, whose regulation involves a large number of binding sites, and some gene promoters with specialized regulatory functions. Most of the regulatory elements are generally present in plant species. Furthermore, this type of analysis can be useful in finding other secondary structures in the 3′ UTR regions of genes or possibly in other sequences. Based on this analysis, we suggest that occurrence of binding sites within a SUS promoter often reflects various ways to analyze and predict all possible transcription factors which are involved in gene regulation and networking. The results are explained along with proposed utilization of these promoters.

Materials and Methods

The objective of this study was to identify and analyze the SUS gene promoter from *Arabidopsis thaliana* isolated through High Throughput Genomic Sequences (HTGS). High Throughput Genomic Sequences is one of the important database used to identify promoter sequences of different genes. The promoter isolated through HTGS approach was further screened for analysis and identification of conserved domains, cis regulatory elements and transcription factor binding sites.

Identification of AGL3 Transcription factors: The subject SUS promoter sequence consists of various transcription factors and their associated transcription factor binding sites. The putative TFBs were identified by using Plant PAN 2.0 TFB analysis software (http://plantpan2.itps.ncku.edu.tw/promoter.php).

Analysis of putative domains: The SUS promoter sequence was further screened for analysis of putative domains through Pfam (https://pfam.xfam.org/) and InterPro (https://www.ebi.ac.uk/interpro/) database. The analysis showed various domains with different functions related to cell proliferation, differentiation and growth.

Alignment of MADS domain: Characterization of SUS sequence revealed various TFBs and conserved domains. One of the most important domain found within AGL3 was MADS DNA binding domain. The comparison and alignment of MADS domain was performed by using Simple modular architecture research tool (SMART) (http://smart.embl-heidelberg.de/smart/ do annotation.pl?ACC =SM00432&BLAST=DUMMY).

Motif analysis of MADS domain: The MADS domain was further screened for analysis of conserved regions and protein signatures within the promoter sequence. This was done by using PRINTS database (http://130.88.97.239/PRINTS/index.php).

Evolutionary existence: To study the evolutionary pathway and emergence of MADS domain, the identified domain was analyzed through database of orthologous organism (http://www.orthodb.org/? level=&species=& query= EOG09370SO1).

Results

Retrieval of sucrose synthase promoter AGL3 transcription factors: The promoter sequence of SUS gene was searched to find putative transcription factor binding sites by using Plant PAN 2.0 TFB analysis software. Figs. (1-3) shows TF binding sequence of AGL3 (AT5G23260), AGL3 (AT2G03710), AGL3 (U81369). Furthermore, results showed three AGL3 associated TF with specific family and domain as describe in table 1 below:

The first reported TF (AT5G23260) plays an important role in developmental regulation of the endothelium and in accumulation of proanthocyanidins (PAs) which are useful for seed pigmentation after oxidation. Moreover, it is also required for normal activation of the BANYULS (*BAN*) promoter in the endothelium body. This promoter is useful to trigger gene expression in *B. napus* seed coat for validation of various candidate genes.

Pfam Domain - SRF-type transcription factor: The query promoter sequence was searched in Pfam and InterPro database. The first domain found was SRF type transcription factor, which functioned as a DNA binding and dimerization domain. SRF is a nuclear protein important for cell proliferation and differentiation. Moreover, this is essentially required for transcriptional regulation of several growth factors. This core domain interacts with certain accessory factors and varies up to 90 amino acids. Within this core, is a DNA binding region referred as MADS box. This region showed similarity to many eukaryotic regulatory proteins. This MADS-box domain is associated with K-box region.

For further verification, the query sequence was analyzed by using InterPro (IPR002487). Further, its two of the functions at molecular level includes DNA binding and Protein dimerization activity. For binding, any gene can interact selectively and non-covalently with DNA. The EMBL-EBI quick go software was used to study the ancestor chart of this DNA binding domain having accession ID (GO 0003677) as shown in fig 4 below:

Table 1. Shows TF found in AGL3 obtained by using PlantPAN 2.0.

Gene name	TF ID	Function	TF binding sequence	TF Family	Species	Pfam Family	Domain
AGL3	a). AT5G23260	Developmental regulation and validation of candidate genes	TTCCAAAAAGGAAA	MADS box; MIKC	Arabidopsis thaliana	PF01486 K-box region	PF00319 SRF-type transcription factor
	b). AT2G03710	Probable transcription factor	TCCATATATAGAA	MADS box; MIKC	Arabidopsis thaliana	PF01486 K-box region	PF00319 SRF-type transcription factor
	c). U81369	Probable transcription factor	TTTTCTATTTTTGGTAA	Others	Arabidopsis thaliana	Not found	Not found

a). Retrieval of TF binding sequence of AGL3 TF (AT5G23260)

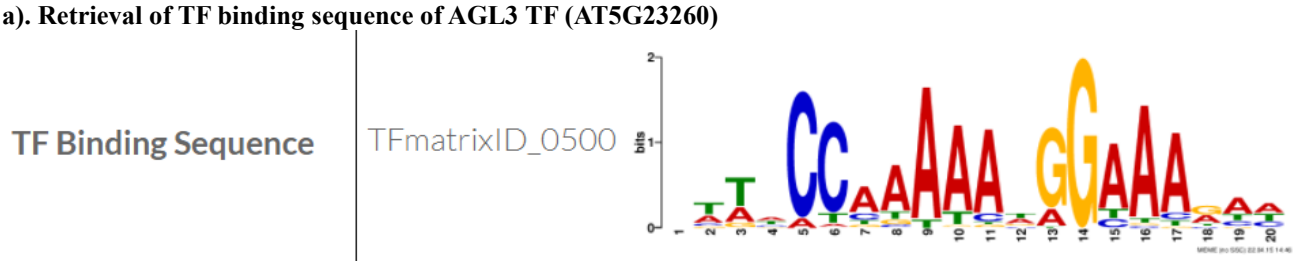

Fig 1. Shows the TF binding sequence of AGL3 (AT5G23260) obtained from PlantPAN 2.0.

b). Retrieval of TF binding sequence of AGL3 TF (AT2G03710)

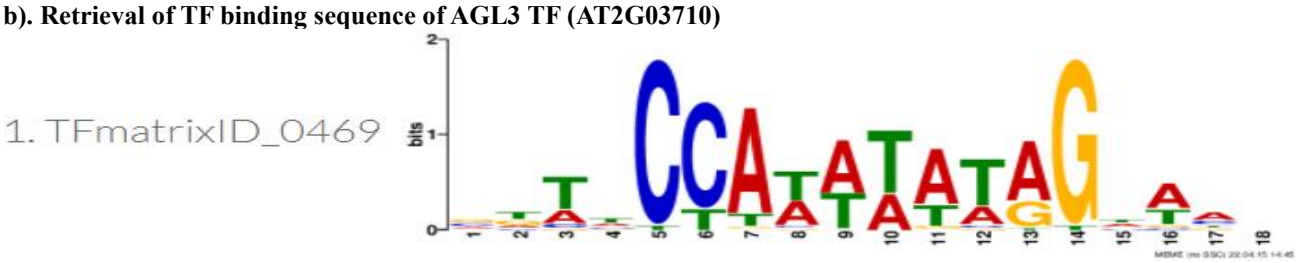

Fig 2. Shows the TF binding sequence of AGL3 (AT2G03710) obtained from PlantPAN 2.0.

c). Retrieval of TF binding sequence of AGL3 TF (U81369)

Fig. 3. Shows the TF binding sequence of AGL3 (U81369) obtained from PlantPAN 2.0.

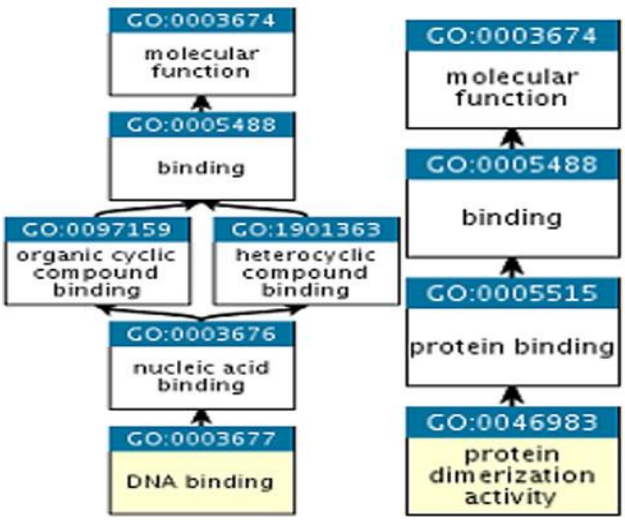

Fig. 4. Shows ancester chart for: a). GO 0003677 (DNA binding), b). GO 0046983 (Protein dimerization activity).

K-box region: The K-box region is a coiled structure and mostly found with SRF-type transcription factors. It plays a vital role in multimer formation.

Alignment of MADS domain by using SMART: The family alignment of MADS domain was obtained by using SMART (Simple modular architecture research tool). The consensus sequence showed 60% similarity. Results are represented diagrammatically in fig 5:

Identification of MADS domain motif: The MADS domain of AGL3 TF was searched for identification of putative motif and protein signatures by using PRINTS database. Fingerprints represent domains. Similarly, MADSDOMAIN is a 3-element fingerprint that provides a signature for the MADS domain. The query sequence showed three motifs in a complete sequence. Results are given in table 2 below:

Evolutionary existance: The MADS box domain was searched in database of orthologous organism (http://www.orthodb.org/?level=&species=&query=EOG09370SO1). The maximum evolutionary rate was approximated. 0.85 as shown in Fig 6. Out of total 2140 genes, 1290 was involved in DNA binding, 1404 had trancription factors with MADS - box domain.

Table 2. Shows the putative number of motif in AGL3 MADS domain.

Protein ID	No of motif	Total No of signatures	Positive	True partial	False partial
MADS-domain fingerprint	3	196	189	7	0

MADS Genes and MIKC structure: MADS genes are commonly known as a key regulator for vegetative and reproductive development. These proteins share a stereotypical MIKC structure. It consists of an N-terminal domain, which is present only in minority of proteins. Some proteins like MADS domains play a major role in DNA binding. Others, keratin like (K-box) domain promotes protein dimerisation activity. However, some of the C- terminal domains are involved in trasncriptional activation.

Discussion

Sucrose commonly distributed compound is used by various plants to remobilize reserves from storage tissues during plant germination and development. It is a major product of photosynthesis and mostly functions in storage and translocation. It also helps for regulation of gene expression (Winter & Huber, 2000; Smeekens, 2000; Wiese *et al*., 2004).

Promoters are short DNA sequences, which occur upstream of the coding regions of genes. They are essential for plant gene expression and regulation (Dare *et al*., 2008). Isolation and identification of dicot promoters helps to understand the transcriptional regulation. This can be performed by analyzing and identifying TF and specific TFBs within the promoter region. The promoter isolated through these studies may be effectively substituted in plant genetic engineering with other constitutive promoter for transgene expression in economically important agricultural crops (Naqvi *et al*., 2017). Moreover, some of the newly designed synthetic promoters have wide range of dispersed motifs that are not as much conserved (Pilpel *et al*., 2001). This can arranged according to need by changing the copy number and adjusting the inner spacing among bases within the promoter sequence (Liu *et al*., 2013; Mehrotra *et al*., 2011; Venter *et al*., 2007; Hammer *et al*., 2006; Rushton *et al*., 2002). Promoters are rich with certain TFs, comprising of DNA binding domains and some effector

domains. Most of the studies showed, presence of DNA binding domain on a-helix of major groove (Pavletich & Pabo, 1991). Variation among these domains was observed due to change in copy number and tandem repeats at particular site (Boch & Bonas, 2010; Bogdanove *et al*., 2010). Alterations in these DNA binding residues of each particular domain can achieved by construction of useful novel synthetic promoters for controlled transcriptional regulation. However, changes in domains will results in change in binding specificity of transcription factors.

Gene expression is a key factor, which regulates the binding of specific proteins called transcription factors with short DNA sequence motifs located on specific binding sites within the promoter region. To determine, the number of cis regulatory elements and transcription factors in model organisms, various bioinformatics tools can be used (Thomas & Chiang, 2006). However, to recognize sequence specific motifs by transcription factors, this is important to identify transcription factor binding site (TFB) first (Carey *et al*., 2001). Therefore, the promoter sequence of sucrose synthase gene was analyzed for identification of TF and TFB's by using computational tools. One of the most common TFB found was AGL3, which is expressed in all above ground vegetative organs, as well as in flowers, but not in roots. It is also important for regulation of transcription. Moreover, some of these are enriched with conserved regions called as domains. These are functional unit of proteins which vary in size from protein to protein. The most commonly found domain in AGL3 TFB was located in MADS-box protein. This is a DNA binding domain, which regulates transcription of several genes involved in structural pathway of AGL3. Furthermore, the AGL3 mostly binds with target sequences of SRF. These results suggest that AGL3 is a widely distributed DNA-binding protein, which may be involved the transcriptional regulation of genes in many cells.

Fig. 5. Shows the MSA results for MADS-Box domain of AGL3 TF by using SMART.

Fig. 6. Shows the evolutionary rate of MADS – box domain obtained by using OrthoDB.

Crystal Structure of MADS domain

The crystal structure of AGL3 MADS domain is represented in Fig. 7 below:

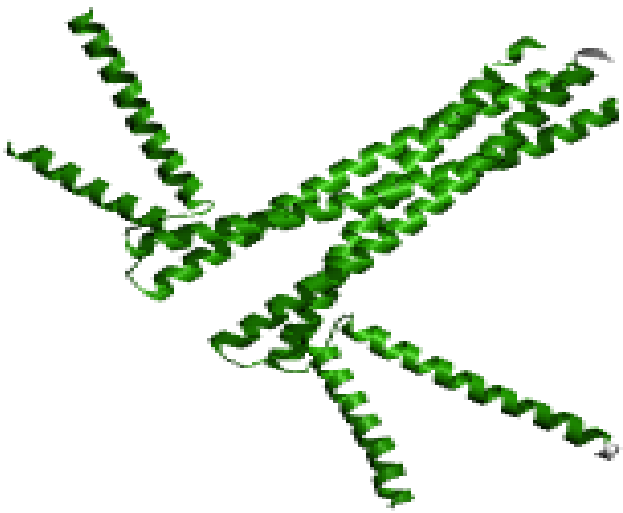

Fig. 7. Shows crystal structure of MADS domain of AGL3 TF.

Conclusion

To understand the gene expression patterns, it is necessary to predict the transcription factor binding sites. Great achievements have been made in identification and characterization of such sites by using high throughput methods. TFB's can be determined by analyzing the associated TF within the promoter region. Moreover, binding of TFBs also depends on some regulatory sequences and conserved protein domains which are essential for understanding gene expression patterns regulated by transcriptional control units.

References

Babu, M. and S.A. Teichmann. 2003. Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucl. Acids Res.*, 31(4): 1234-1244.

Baud, S., M.N. Vaultier and C. Rochat. 2004. Structure and expression profile of the sucrose synthase multigene family in Arabidopsis. *J. Exp. Bot.*, 55(396): 397-409.

Boch, J. and U. Bonas. 2010. Xanthomonas AvrBs3 family-type III effectors: discovery and function. *Ann. Review Phytopathol.*, 48.

Bogdanove, A.J., S. Schornack and T. Lahaye. 2010. TAL effectors: finding plant genes for disease and defense. *Curr. Opin. in Plant Biol.*, 13(4): 394-401.

Carey, M., C.L. Peterson and S.T. Smale. 2009. *Transcriptional regulation in eukaryotes*. Cold Spring Harbor Laboratory Press.

Dare, A.P., R.J. Schaffer, K. Lin-Wang, A.C. Allan and R.P. Hellens. 2008. Identification of a cis-regulatory element by transient analysis of co-ordinately regulated genes. *Plant Methods*, 4(1): 17.

Fallahi, H., G.N. Scofield, M.R. Badger, W.S. Chow, R.T. Furbank and Y.L. Ruan. 2008. Localization of sucrose synthase in developing seed and siliques of Arabidopsis thaliana reveals diverse roles for SUS during development. *J. Exp. Bot.*, 59(12): 3283-3295.

Hammer, K., I. Mijakovic and P.R. Jensen. 2006. Synthetic promoter libraries–tuning of gene expression. *Trends in Biotechnol.*, 24(2): 53-55.

Liu, W., J.S. Yuan C.N. Jr. Stewart. 2013. Advanced genetic tools for plant biotechnology. *Nature Rev. Gen.*, 14(11): 781.

Martínez-Antonio, A., S.C. Janga, H. Salgado and J. Collado-Vides. 2002. Internal-sensing machinery directs the activity of the regulatory network in Escherichia coli. *Trends in Microbiol.*, 14(1): 22-27.

Masood, A., N. Iqbal, H. Mubeen, R.Z. Naqvi, A. Khatoon and A. Bashir. 2017. Cloning and expression analysis of d-hordein hybrid promoter isolated from barley (*Hordeum vulgare* L.). *Pak. J. Bot*, 49(3): 1085-1095.

Mehrotra, R., G. Gupta, R. Sethi, P. Bhalothia, N. Kumar and S. Mehrotra. 2011. Designer promoter: an artwork of cis engineering. *Plant Mol. Biol.*, 75(6): 527-536.

Messenguy, F. and E. Dubois. 2003. Role of MADS box proteins and their cofactors in combinatorial control of gene expression and cell development. *Gene*, 316: 1-21.

Naqvi, R.Z., H. Mubeen, A. Masood, A. Khatoon and A. Bashir. 2017. Identification, isolation and evaluation of a constitutive sucrose phosphate synthase gene promoter from tomato. *Pak. J. Bot*, 49(3): 1105-1112.

Pavletich, N.P. and C.O. Pabo. 1991. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science*, 252(5007): 809-817.

Pilpel, Y., P. Sudarsanam and G.M. Church. 2001. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Gen.*, 29(2): 153.

Potenza, C., L. Aleman and C. Sengupta-Gopalan. 2004. Targeting transgene expression in research, agricultural, and environmental applications: promoters used in plant transformation. *In Vitro Cell Dev. Biol. Plant*, 40(1): 1-22.

Ptashne, M. and A. Gann. 2002. *Genes and signals*. Cold Spring Harbor Laboratory Press.

Rushton, P.J., A. Reinstädler, V. Lipka, B. Lippok and I.E. Somssich. 2002. Synthetic plant promoters containing defined regulatory elements provide novel insights into pathogen-and wound-induced signaling. *The Plant Cell*, 14(4): 749-762.

Seshasayee, A.S., P. Bertone, G.M. Fraser and N.M. Luscombe. 2006. Transcriptional regulatory networks in bacteria: from input signals to output responses. *Curr. Opin. in Microbiol.*, 9(5): 511-519.

Smeekens, S. 2000. Sugar-induced signal transduction in plants. *Ann. Rev. Plant Biol.*, 51(1): 49-81.

Stormo, G.D. 2000. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1): 16-23.

Thomas, M.C. and C.M. Chiang. 2006. The general transcription machinery and general cofactors. *Critical Rev. in Biochem. & Mol. Biol.*, 41(3): 105-178.

Venter, M. 2007. Synthetic promoters: genetic control through cis engineering. *Trends in Plant Sci.*, 12(3): 118-124.

Wiese, A., N. Elzinga, B. Wobbes and S. Smeekens. 2004. A conserved upstream open reading frame mediates sucrose-induced repression of translation. *The Plant Cell*, 16(7): 1717-1729.

Winter, H. and S.C. Huber. 2000. Regulation of sucrose metabolism in higher plants: localization and regulation of activity of key enzymes. *Critical Rev. Plant Sci.*, 19(1): 31-67.

Yada, T., M. Nakao, Y. Totoki and K. Nakai. 1999. Modeling and predicting transcriptional units of Escherichia coli genes using hidden Markov models. *Bioinformatics*, 15(12): 987-993.

(Received for publication 27 March 2018)