African Journal of Biotechnology Vol. 8 (6), pp. 1121-1126, 20 March, 2009
Available online at http://www.academicjournals.org/AJB
ISSN 1684–5315 © 2009 Academic Journals

Full Length Research Paper

RNA recognition motif (RRM)-containing proteins in *Bombyx mori*

Lin-Ling Wang and Ze-Yang Zhou*

College of Life Sciences, Chongqing Normal University, Chongqing 400047, China.

Accepted 28 January, 2009

RNA recognition motif (RRM)-containing proteins play important roles in the processing of RNA and regulation of protein synthesis. *Bombyx mori* is a model species of Lepidoptera whose genome sequence is available; however, its RRM-containing proteins have not been thoroughly analyzed. In this study, 134 RRM-containing protein genes in Drosophila melanogaster were used to perform BLAST search against *B. mori* genome- and EST databases. Results showed that 123 genes have orthologues in the *B. mori* genome database and 89 in *B. mori* EST database, indicating some of the genes were not expressed in *B. mori*. 23 RRM-containing protein genes that have complete open reading frame were obtained through contig assembly in *B. mori* EST database and then registered in NCBI. Cluster analysis showed that these 23 new proteins formed two distinct clades. Especially, TAR DNA-binding protein-43(TBPH) gene of *B. mori* is different from that of other species in that it does not have introns. Moreover, the length of TBPH of *B. mori* is much shorter than that of other species' TBPH, suggesting that it may play some roles in emergency responses.

Key words: *Bombyx mori*, RRM-containing protein, TAR DNA-binding protein-43.

INTRODUCTION

RNA binding proteins constitute an extraordinarily complex class of cellular factors, and there are many types of RNA interaction and more are to be discovered. The RNA Recognition Motif (RRM), sometimes referred to as RNP1, is one of the first identified domains for RNA interaction. RRM is very common across species, which suggests it may be an ancient fold (Bandziulis et al., 1989; Varani and Nagai, 1998). It is composed of about 80 amino acids that fold in a βαββαβ structure arranged as a four-stranded β-sheet packed against the two α helices. A few, highly conserved, aromatic amino acids within the β-sheet constitute the RNP1 and RNP2 motifs and contribute to stacking interactions with the RNA nucleotides, while residues in the loops between α-helices and β-strands confer sequence recognition features (Varani and Nagai, 1998). This domain is typical of many splicing factors and hnRNP proteins whose multiple domains often contribute to specific RNA binding. Proteins with RRMs are involved in a large number of processes through specific interactions with RNA (Dreyfuss et al., 2002), but a subset of RRMs can also interact with other proteins (Kielkopf et al., 2004).

In *Drosophila*, the dipteran model, RRM-containing proteins have been broadly studied. *Drosophila melanogaster* genome encodes over 100 RRM-containing proteins (Lasko, 2000; Gamberi et al., 2006), including splicing factors and components of the 3' end processing machinery (Lasko, 2000; Salz et al., 2004). In *Bombyx mori*, the lepidoptera model, few of its RRM-containing proteins have been studied.

In this study, 134 RRM-containing protein genes in *Drosophila* were used to perform BLAST search against *B. mori* genome- and EST database. This study provides the first systematic bioinformatics analysis of RRM-containing proteins in *B. mori* and may serve as a basis for the characterization of these proteins.

MATERIALS AND METHODS

Identification of RRM-containing protein genes in *B. mori* genome- and EST databasesKnown RRM-containing protein genes in *Drosophila* were used to perform BLAST search against *B. mori* genome- and EST databases, E values <0.001 required. The assignments were confirmed by manually examining protein lengths

*Corresponding author. E-mail: zyzhou@cqnu.edu.cn. Tel.: +86-23- 65910315; Fax: +86-23- 65910312.

and domain structures, and then orthologous proteins were assigned with similar names. All identified proteins were examined in the non-redundant protein database at the National Center for Biotechnology Information (NCBI) using BLASTp.

Conserved domain analysis

Conserved domains were analyzed at (<http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>) by Search against database (CDD-24291PSSMs) with Reverse Position Specific BLAST (Marchler-Bauer et al., 2007).

Cloning of five of the newly assembled RRM-containing protein genes

Total RNA was extracted using Rneasy Mini Kit of QIAGEN (Hilden, Germany) and treated with RNase-free DNase I. RT-PCR was carried out following the manufacturer's instructions. All primers were designed based on the EST sequences of silkworm RRM-containing proteins. Primers Aly-sense (5'TAT AAA AGC CAA CAA AAA GAC AAG3'), Aly-antisense (5'TGC AGT ACC TAA AGA TCT ACC TGA3'), eIF3S4-sense (5'CCT ATC AAT TAT CAC TTC ATA CGA TA3'), eIF3S4-antisense (5'TAT TTA AGT TTG TGT ACA AAT GGA AC3'), CstF-64-sense (5'TTT ACA ATT CAA CGC ATA AAA A3'), CstF-64-antisense (5' CGC ATG TCC TGG TCG AGC AGC 3'), eIF3-S9-sense (5'ATC GTC AAA ATG GCT AAG AAA A3'), eIF3-S9-antisense (5'GCG TAG CGT ATA GCG ACT ACT C3'), TBPH-sense (5'GAG ATG AGG AAG TGG AAG GTC A3') and TBPH-antisense (5'CAA TGT AGG TTT GTG GAG AAC A3') were used for the amplification of Aly, CG10881-like protein, CstF-64, eIF3-S9 and TBPH genes of *B. mori*. PCR was carried out in a 25-µl reaction system containing normalized cDNA, 1.5 mM $MgCl_2$, 50 mM KCl, 10 mM Tris–HCl (pH 8.3), 0.25 mM each dNTP, and 2.5 units of Taq DNA polymerase.

The PCR conditions for the amplification of cDNA fragments were 1 cycle at 94°C for 2 min; 34 cycles at 94°C for 30 s, 52°C for 2 min, and 72°C for 40 s; and 1 cycle at 72°C for 10 min. PCR products were separated by gel electrophoresis, recovered from the gel using the 3S spin DNA gel purification kit V3.1 (Shenergy Biocolor, Shanghai, China) and TA cloned using PMD 18-T vector (TaKaRa, Dalian, China). Plasmid was prepared using Go3S spin plasmid miniprep kit V3.1 (Shenergy Biocolor, Shanghai, China) and sequenced using CEQ 8000 genetic analysis system (Beckman Coulter), all operation performed according to manufacturer's instructions.

Phylogenetic analysis

The 23 newly assembled RRM-containing protein genes were aligned using CLUSTALX version 1.8 (Thompson et al., 1997). Neighbor-joining tree was constructed using MEGA 3.0 (Kumar et al., 2004) under the two-parameter distance of Kimura (Kimura, 1980).

RESULTS

Overview of RRM-containing proteins in *B. mori*

134 RRM-containing protein genes in *Drosophila* (Gamberi et al., 2006) were used to perform BLAST search against *B. mori* genome- and EST database, and 123 genes were founded to have orthologues in the *B. mori* genome database and 89 in *B. mori* EST database. Approximately 54% of *B. mori* RRM proteins contain more than one RRM domains. In addition to RRM domains, there are many other conserved domains in RRM-containing proteins, including ZnF-RBZ, ZnF-C3H1, G-patch, LA, tRNA-U5-meth-tr and rADc. Poly (A) binding protein (Pabp) has four RRM motifs and one PolyA binding conserved domain, which is consistent with its predicted roles as positive regulator of translation and Poly (A) binding protein. Cabeza (Caz) has one RRM and one ZnF-RBZ conserved domain which binds RanGDP. RNA binding protein 5 (RBM5) has one ZnF-RBZ, one G-patch and two RRM domains. With seven highly conserved glycines in it, G-patch domain is found in a number of RNA binding proteins, suggesting that this domain may have a RNA binding function. In *Drosophila*, it was involved in apoptosis, cell proliferation, histone mRNA 3'end processing and nuclear mRNA splicing.

Some RRM-containing protein genes in *Drosophila* have no orthologue in *B. mori* EST database, but have orthologue in *B. mori* genome database. For example, the gene encodes Bruno which is a translational repressor and spliceosomal snRNP protein in *Drosophila*. Whether these genes were expressed in *B. mori* remains to be studied. Of the 89 genes expressed in *B. mori*, ~99% have not yet been characterized or registered in NCBI. Molecular functions of some RRM proteins just can be predicted based on their orthologous proteins in *Drosophila*, yeast and human (Gamberi et al., 2006). There are 11 RRM-containing protein genes in *Drosophila* which do not have orthologues in the *B. mori* genome- and EST databases.

23 newly assembled RRM-containing proteins having complete open reading frame in *B. mori*

23 newly assembled RRM-containing protein genes having completely open reading frame in *B. mori* are listed Table 1 and then registered in NCBI. These proteins have similar length and high homology in conserved domain with their orthologous protein in *Drosophila*. In order to verify these genes, with *B. mori* cDNAs as templates, 5 of the 23 genes, namely, Aly, eIF3S4, CstF-64, eIF3-S9 and TAR DNA-binding protein-43 (TBPH), were PCR-amplified, cloned and sequenced. Results suggest that our EST assembling is correct. For example, ALY gene, consisting of four exons and three introns, lies in Ctg006950 (AADK01006950), and just has one copy in the *B. mori* genome. ALY is composed of 254 amino acids with one RRM motif between 101 and 173 amino acids. *B. mori* ALY is 62% (Expect = 2e-52), 65% (Expect = 6e-64), 58% (Expect = 3e-60) identical in amino acid sequence to *D. melanogaster*, *Tribolium castaneum* and *Anopheles gambiae str. PEST* ALY, respectively. And take another example, TBPH gene has no introns, and its products consists of 286 amino acid residues, with the two RRM motifs located between 90^{th} and 160^{th} as well as between 174^{th} and 223^{rd} amino acid residues. TBPH of *B.*

Table 1. 23 newly assembled RRM-containing proteins in *Bombyx mori*.

Protein	Conserved Domain	Copies in genome	Predicted molecular function (Gamberi, 2006)	Accession number in Genbank
ALY THO complex subunit 4	1RRM	1	Transcriptional coactivation Nucleic acids transport RNA nucleocytoplasmic export	DQ497195
Cap binding protein 20 (Cbp20)	1RRM	1	Nuclear mRNA splicing Putative involvement in nucleocytoplasmic transport	DQ497196
Eukaryotic translation initiation factor 3 subunit 4 (eIF3S4)	1RRM	1	Translation initiation eIF3 complex	DQ497197
Basal transcriptional activator hABT1 (ABT1)	1RRM	1	--	DQ497198
Pre-mRNA branch site protein p14 (SF3B14)	1RRM	1	--	DQ497199
CG15440-like protein SECP43 protein	2RRM	1	Nuclear mRNA splicing Possible DNA binding	DQ497200
eIF3-S9	1RRM 1Trp-Asp (WD) repeat domain	1	Translation initiation	DQ497202
RNA binding protein 1 (Rbp1)	1RRM	1	Nuclear mRNA splicing	DQ497203
Repressor splicing factor 1 (Rsf1)	1RRM	1	Negative regulation of nuclear mRNA splicing	DQ497204
the TAR DNA-binding protein-43 (TBPH)	2RRM	1	Nuclear mRNA splicing	DQ497205
Tsunagi (Tsu) RNA binding protein 8A/Y14 (RBM8A)	1RRM	1	mRNA localization Embryonic axes determination	DQ497206
xl6 (xl6) Splicing factor, arginine/serine-rich 7 (SFRS7)	1RRM	1	Nuclear mRNA splicing	DQ497207
Poly(A) binding protein (PAbp)	4RRM 1PolyA	2	Positive regulator of translation Poly(A) binding	DQ646407
Nucleolysin TIAR	3RRM 1Calcipressin		Nuclear mRNA splicing U1 SnRNP	DQ646410
eIF4B	1RRM		Translation initiation eIF4E binding	DQ646408
B52 Splicing factor, arginine/serine-rich 6 (SFRS6)	2RRM	1	Nuclear mRNA splicing Possible role in chromatin condensation	DQ648521
CG10466 RNA binding motif protein, X-linked 2 (RBMX2)	1RRM	1	mRNA processing Proteolysis and peptidolysis Zn binding	DQ648522
Splicing factor 45 (RBM17)	1RRM	1	Nuclear mRNA splicing	DQ648523
Polymerase delta interacting protein 3 (POLDIP3)	1RRM	1	--	DQ648524

Table 1. Contd.

Nucleolar phosphoprotein Nopp34	1RRM	1	--	DQ648525
Splicing factor, praline- and glutamine-rich (SFPQ)	2RRM	1	--	DQ648526
Pabp2 Polyadenylate binding protein 2 (PABN1)	1RRM	1	mRNA polyadenylation Poly(A) binding	DQ648527
Spliceosomal protein on the X (Spx)	2RRM	1	Nuclear mRNA splicing U2 snRNP	DQ648528

mori is 58% (Expect = 7e-88), 60% (Expect = 7e-89), 52% (Expect = 5e-81) identical in amino acid sequence to that of *D. melanogaster*, *T. castaneum* and *A. gambiae str. PEST* TBPH, respectively.

Conserved sequence analysis of *B. mori* RRM-containing proteins showed that these 23 RRM-containing proteins have at least one RRM motif, with some having more than two RRM motifs, including TBPH, Poly (A) binding protein, Nucleolysin TIAR; arginine/ serine-rich 6 (SFRS6) and spliceosomal protein on the X (Spx). Apart from the RRM motif, eIF3-S9 has a Trp-Asp (WD) repeat domain, Poly (A) binding protein (PAbp) has a Poly (A) binding domain, and Nucleolysin TIAR has a Calcipressin domain.

In order to reveal functional relationships, multiple sequence alignments were performed, and a phylogeneric tree was constructed of the 23 newly assembled RRM proteins in *B. mori* (Figure 1). Phylogenetic tree showed that some proteins with similar functions were grouped together.

These 23 new RRM proteins in *B. mori* formed two distinct clades. The first clade consists of Cbp20, ALY, TBPH, Tsu, etc, which are predicted to be involved in nuclear mRNA splicing, RNA nucleocyto-plasmic export, mRNA localization and embryonic axes determination. The secondary clade consists of Spx, Rbp1, xl6, etc, most of which are predicted to be involved in nuclear mRNA splicing, Poly (A) binding and translation regulation, with some having no functional information.

DISCUSSION

A total of 123 RRM-containing proteins was found in *B. mori*, 134 in *Drosophila* (Gamberi et al., 2006), 196 in *Arabidopsis thaliana* (Lorković and Barta, 2002), 43 in *Ustilago maydis* (Becht et al., 2005), 139 in *Trypanosoma cruzi*, 75 in *Trypanosoma brucei*, and 80 in *Leishmania major* (Gaudenzi et al., 2005). Clearly, organisms express a complex set of RRM-containing proteins, many of which are present in all higher eukaryotes (Lopato et al., 1996; Lorković and Barta, 2002; Gamberi et al., 2006).

Some of RRM-containing proteins found in *B. mori* can be grouped according to their possible functions. There are 35 proteins predicted to be involved in splicing, 5 in mRNA processing, 13 in translation repression, regulation or initiation, 8 in transcription, and 3 in apoptosis. For example, Cap Binding Protein 20 (Cbp20) is the small subunit of the nuclear Cap Binding Complex involved in splicing and nucleocytoplasmic transport (Izaurralde et al., 1994, 1995); xl6, also called as splicing factor, arginine/serine - rich 7 (SFRS7), consists of many arginine and serine repetitions on the C terminal and is predicted to be involved in splicing (Allemand et al., 2001, 2002); X-linked 2 (RBMX2), a RNA binding motif protein, is predicted to be involved in mRNA processing, peptidolysis and Zn binding (Kaminker et al., 2002); eIF4B, eIF3S4 and eIF3-S9 are translation initiation factors; and Nucleolysin TIA-1 (TIA1) and RNA binding protein 5 (RBM5) are predicted to be involved in apoptosis (Lasko, 2000; Anholt and Mackay, 2001). It should be noted that many RNA-binding proteins are in fact involved at multiple stages of RNA processing, transport, and degradation.

There are 11 RRM-containing protein genes in *Drosophila* which do not have orthologues in *B. mori* genome. Some of the 11 genes do not have orthologues in other species, either. For example, except the genes encoding CG12288 and modulo (Mod), nine of the 11 genes do not orthologues in human genome. Moreover, seven of the 11 RRM-containing proteins do not have functional information in *Drosophila*, suggesting that these genes might have redundant functions and were gradually eliminated in evolution.

The majority of RRM-containing protein genes consist of multiple exons and introns, with a few having no introns such as TBPH of *B. mori*. TBPH of *B. mori*, consisting of 286 amino acid residues and having two RRM motifs, shows high homology with that of other species within conserved domains, but differs greatly in length from that of other species, with TBPH of *Macaca mulatta*, *Homo sapiens*, *Musca domestica* and *Mus musculus* having 413 amino acid residues, that of *D. melanogaster* 332, and that of *Xenopus laevis* 245 (UniGene-Hs.709233) http://www.ncbi.nlm.nih.gov/UniGene/clust.cgi?ORG=Hs&

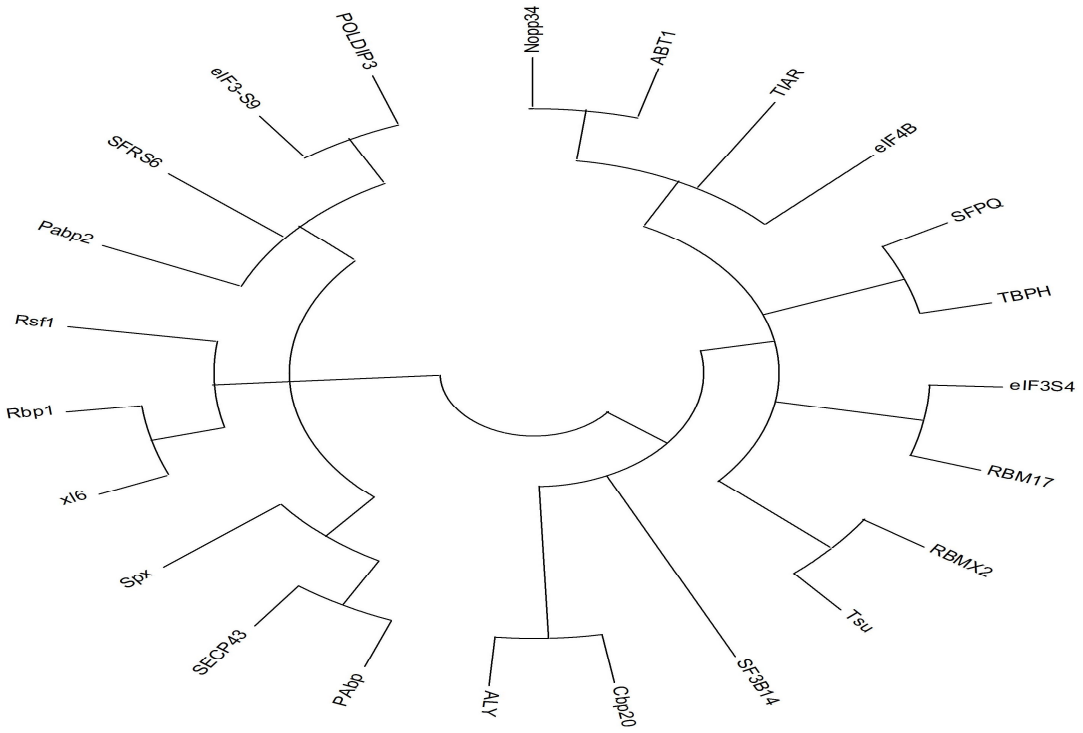

Figure 1. Phylogenetic tree of the 23 newly assembled RRM-containing proteins. Alignments of sequences were performed with the ClustalX Neighbor-joining tree was constructed using MEGA 3.0.

CID=709233). The difference between TBPH of *B. mori* and that of other species lies mainly in that the latter has more amino acid residues at C terminal. The C terminal of other species was analyzed for conserved domains, and no conserved domains were found, suggesting that TBPH of *B. mori* minimizes its length while maintaining the conserved domains of TBPH.

There are 10 pieces of data on TBPH gene in *B. mori* EST database, with 1 piece of data derived from wing disk, 2 from verson's gland, 1 from compound eye, 1 from diapause-destined embryo, 3 from ovary, and the remaining 2 from unspecifie dtissues, The expression of BPH gene in so many tissues indicates that TBPH may be a house-keeping gene that plays important roles.As mentioned above, TBPH gene of *B. mori* has no introns. However, TBPH gene of *D. melanogaster* consists of 7 exons (BLAST Assembled Genomes). It is worth studying that why TBPH gene of *B. mori* has no introns despite that fact that the genome of *B. mori* is much larger than that of *D. melanogaster*. In eukaryotes, genes that do not have introns account for a small portion of genes, and generally, they are functionally conserved regulatory factors which are transcribed and translated fast in emergency situations. Take hsp70 as an example: hsp70, a widespread and conserved protein, can be divided into two classes: i.) hsp70 with introns that is expressed constitutively, also called as hsc70; and ii.) hsp70 without introns that is expressed in emergency situations and is transcribed and translated fast (Daugaard et al., 2007).

The fact that TBPH gene of *B. mori* has no introns and that the length of TBPH of *B. mori* is much shorter than that of other species suggest that TBPH of *B. mori* may play some roles in emergency responses other than nuclear mRNA splicing.

ACKNOWLEDGMENTS

This study was supported by the National Basic Research Program of China under grant (No. 2005CB121000), the Science and Technology Research Program of Chongqing municipal education commission (No.KJ080815).

REFERENCES

Allemand E, Dokudovskaya S, Bordonne R, Tazi J (2002). A conserved *Drosophila* transportin-serine/arginine-rich (SR) protein permits nuclear import of *Drosophila* SR protein splicing factors and their antagonist repressor splicing factor 1. Mol Biol Cell. 13(7): 2436-2447.

Allemand E, Gattoni R, Bourbon HM, Stevenin J, Cáceres JF, Soret J, Tazi J (2001). Distinctive features of *Drosophila* alternative splicing factor RS domain: implication for specific phosphorylation, shuttling, and splicing activation. Mol Cell Biol. 21(4): 1345-1359.

Anholt RR, Mackay TF (2001). The genetic architecture of odor-guided behavior in *Drosophila melanogaster*. Behav Genet. 31(1): 17-27.

Bandziulis RJ, Swanson MS, Dreyfuss G (1989). RNA-binding proteins as developmental regulators. Genes Dev. 3: 431-437.

Becht P, Vollmeister E, Feldbrugge M (2005). Role for RNA-Binding Proteins Implicated in Pathogenic Development of *Ustilago maydis*. Eukaryot Cell. 4(1): 121-133.
Dreyfuss G, Kim VN, Kataoka N (2002). Messenger-RNA-binding proteins and the messages they carry. Nat Rev Cell Biol. 3: 195-205.
Gamberi C, Johnstone O, Lasko P (2006). *Drosophila* RNA Binding Proteins. Int Rev Cytol. 248: 43-139.
Gaudenzi JD, Frasch AC, Clayton C (2005). RNA-Binding Domain Proteins in Kinetoplastids: a Comparative Analysis. Eukaryot Cell. 4(12): 2106-2114.
Izaurralde E, Lewis J, Gamberi C, Jarmolowski A, McGuigan C, Mattaj IW (1995). A cap-binding protein complex mediating U snRNA export. Nature. 376: 709-712.
Izaurralde E, Lewis J, McGuigan C, Jankowska M, Darzynkiewicz E, Mattaj IW (1994). A nuclear cap binding protein complex involved in pre-mRNA *splicing*. Cell. 78: 657-668.
Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, Patel S, Frise E, Wheeler DA, Lewis SE, Rubin GM, Ashburner M, Celniker SE (2002). The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. Genome Biol. 3(12): Research 0084.
Kielkopf CL, Lücke S, Green MR (2004). U2AF homology motifs: protein recognition in the RRM world. Genes Dev. 18: 1513-1526.
Kimura M (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16: 111-120.
Kumar S, Tamura K, Nei M (2004). MEGA3: Integrated Software for Molecular Evolutionary Genetics Analysis and Sequence Alignment. Brief Bioinform. 5: 150-163.
Lasko P (2000). The *Drosophila melanogaster* genome: Translation factors and RNA binding proteins. J Cell Biol. 150: F51-F56.
Lopato S, Waigmann E, Barta A (1996). Characterisation of a novel arginine/serine-rich splicing factor in *Arabidopsis*. Plant Cell. 8: 2255-2264.
Lorković ZJ, Barta A (2002). Genome analysis: RNA recognition motif (RRM) and K homology (KH) domain RNA-binding proteins from the flowering plant *Arabidopsis thaliana*. Nucleic Acids Res. 30(3): 623-635.
Daugaard M, Rohde M, Jäättelä M (2007). The heat shock protein 70 family: Highly homologous proteins with overlapping and distinct functions, FEBS Lett. 581: 3702-3710.
Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, Hao L, He S, Hurwitz DI, Jackson JD, Ke Z, Krylov D, Lanczycki CJ, Liebert CA, Liu C, Lu F, Lu S, Marchler GH, Mullokandov M, Song JS, Thanki N, Yamashita RA, Yin JJ, Zhang D, Bryant SH (2007). CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res. 35(D): 237-240.
Salz HK, Mancebo RS, Nagengast AA, Speck O, Psotka M, Mount SM (2004). The Drosophila U1-70K protein is required for viability, but its arginine-rich domain is dispensable. Genetics, 168: 2059-2065.
Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997). The CLUSTAL-X windows interference: Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25: 4876-4882.
Varani G, Nagai K (1998). RNA recognition by RNP proteins during RNA processing. Annu Rev Biophys Biomol Struct. 27: 407-445.