From the Research Laboratories

COMPUTATIONAL ANALYSIS OF THE MESSENGER RNA VARIANTS ENCODING TWO ISOFORMS OF THE HIGH-MOBILITY GROUP BOX 1 PROTEIN

[1)]Luchezar Karagyozov, [2)]Jordana Todorova

[1)]Sofia University "St. Kliment Ohridski" (Bulgaria)

[2)] "Roumen Tsanev" Institute of Molecular Biology – BAS (Bulgaria)

Abstract. High-mobility group box 1 protein (HMGB1) is a multifunctional nonhistone chromosomal protein. This widespread nuclear protein has a dual function-in the nucleus - it binds DNA and participates in practically all DNA-dependent processes. On the other hand, the protein plays an important role in the extracellular matrix as an "alarmin", which interacts with certain receptors and stimulates biochemical pathways, associated with carcinogenesis and metastasis. HMGB1 is a critical damage-associated molecular pattern molecule, has been implicated in several inflammatory diseases and cancer types.

This universality makes it an attractive target for innovative therapeutic strategies in the treatment of various diseases.

The updated database for the *HMGB1* gene, encoding the high-mobility group box 1 protein, was used for computational analysis of the annotated mRNA splice variants. Results showed that five of the splice variants encode an HMGB1 protein, containing 215 amino acid residues. However, two of the splice variants encode a shorter HMGB1 protein with 158 residues. Presently, the existence of a shorter HMGB1 protein is not registered in the protein databanks. This inconsistency is not yet resolved.

Keywords: high-mobility group box 1 protein; HMGB1; splicing; translation; isoforms

Abbreviations

NCBI – National Center for Biotechnology Information, Bethesda, MD
bp – base pairs
kb – kilobase pairs
mRNA – messenger RNA
AUG – a triplet coding methionine. Start of translation.
UAA, UAG – triplets coding no amino acids. Stop codons.

Introduction

The high-mobility group box 1 (HMGB1) is a multifunctional nuclear non-histone DNA binding protein. It is an abundant nuclear protein that has a dual function in the nucleus, it binds DNA and participates in practically all DNA-dependent processes serving as an architectural factor. Outside the cell, HMGB1 plays a different role—it acts as an alarmine that activates a large number of HMGB1-"competent" cells and mediates a broad range of physiological and pathological responses (Bianchi et al. 2005; Ugrinova & Pasheva 2017).

The interest in this protein sharply increased in the early 1990s with the discovery of a new kind of structural domain involved in the interactions with DNA (Jantzen et al. 1990), which displayed similarity to two homologous repeats of an 80-amino acid sequence in HMGB1 (Bianchi et al. 1992; Ugrinova & Pasheva 2017). The protein participates in the spatial organization of the DNA in chromosomes and thus, takes part in the regulation of transcription This protein may be secreted out of the cell in response to stimuli or may be released by damaged cells. HMGB1 plays a role in such diverse processes as inflammation and the spread of cancer. For further information about the gene and the protein visit Online Mendelian Inheritance in Man, entry 163905[1)] (Amberger et al. 2019).

The human *HMGB1* gene maps on chromosome 13 (map position 13q12.3) (Ferrari et al. 1996). The gene is large, it spans 161 kb. As in other eukaryotic protein-coding genes, the information on the *HMGB1* gene is discontinuous. The gene is split into alternating regions with different functions: exons and introns. The sequence of the exons is present also in the mature mRNA, the intron sequences separate the exons and are left out of the functional messenger.

When the HMGB1 gene is active, the exons and the intron are transcribed (copied) into RNA by the enzyme RNA polymerase. The primary transcript is a pre-mRNA, a precursor of the messenger. The pre-mRNA cannot function as a template for protein synthesis. To be translated the pre-mRNA should be processed, the introns should be removed and the exons stitched together. This process is termed splicing *(*Lodish et al. 2008).

The HMGB1 gene contains multiple introns as do most of the protein-coding genes. With many intron/exon junctions, the splicing may proceed in a variety of ways and some exon may be skipped. The alternative splicing may produce several mRNA variants, which are translated into related but different proteins (or isoforms).

The database related to the human HMGB1 gene at NCBI was updated in August 2020; the previous annotation was in 2013 (release 105). Thus, we aimed to analyze the deposited information, to construct a graphic view of the

complex intron-exon combinations presented by the mRNA variants, and to examine the translation products.

Materials and Methods

The entry in the NCBI gene database[2)] for the human high-mobility group box 1 protein (HMGB1, GeneID: 3146) was used to search for basic information. The size and the accession number for the mRNA variants are: variant (1) 5030 bp, NM_001313893; variant (2) 5456 bp, NM_002128; variant (3) 5482 bp, NM_001313892; variant (4) 5506 bp, NM_001363661; variant (5) 5611 bp, NM_001370339; variant (6) 5687 bp, NM_001370340; variant (7) 5483 bp, NM_001370341.

The different regions in the HMGB1 protein are described in detail in the UniProt entry P09409[3)]. Clone Manager Suite 8 (Scientific and Educational Software[4)] was used to align, compare, and analyze the nucleotide and amino acid sequences.

Results and discussion.

The HMGB1 exons are a fraction of the gene

In the human genome, a single gene encodes the high mobility group box 1 protein. The HMGB1 gene is transcribed as a single transcript. The complete transcript (the primary transcript) is 160,893 nucleotide long (see the NCBI gene entry 3146).

On the other hand, the NCBI dataset lists seven HMGB1 mRNA variants, which are similar in length (4830 – 5687 bp) (see Materials and Methods). Each variant is composed of a characteristic set of five or six exons (Table 1). There are twelve different exons in the mRNAs; the combined length of their sequences is 6748 bp. Thus, the sizes of the primary transcript (161 kb) and the exons (7 kb) are vastly different. This shows that in the course of the mRNA maturation 154 kb of nucleotides of the primary transcript (the pre-mRNA) are excised and discarded.

Table 1. The exon composition of the *HMGB1* mRNA variants.

Each mRNA variant represents a unique combination of five (or six) exons. The exons in the primary transcript are numbered according to their order (5' to 3') (the left-most column). The length of each exon and its position in the mRNA variants are indicated.

Exon	Length (bp)	mRNA var. (1) Exon range:	mRNA var. (2) Exon range:	mRNA var. (3) Exon range:	mRNA var. (4) Exon range:	mRNA var. (5) Exon range:	mRNA var. (6) Exon range:	mRNA var. (7) Exon range:
1	927	exon 1/5 1-927						
2	372						exon 1/5 1-372	
3	141		exon 1/5 1-141		exon 1/6 1-141	exon 1/5 1-141		
4	168							exon 1/5 1-168
5	167			exon 1/5 1-167				
6	164	exon 2/5 928-1091	exon 2/5 142-305	exon 2/5 168-331	exon 2/6 142-305	exon 2/5 142-305	exon 2/5 373-536	exon 2/5 169-332
7	146	exon 3/5 1092-1237	exon 3/5 306-451	exon 3/5 332-477	exon 3/6 306-451	exon 3/5 306-451	exon 3/5 537-682	exon 3/5 333-478
8	175	exon 4/5 1238-1412	exon 4/5 452-626	exon 4/5 478-652	exon 4/6 452-626	exon 4/5 452-626	exon 4/5 683-857	exon 4/5 479-653
9	50				exon 5/6 627-676			
10	4985					exon 5/5 627-5611		
11	3618	exon 5/5 1413-5030						
12	4830		exon 5/5 627-5456	exon 5/5 653-5482	exon 6/6 677-5506		exon 5/5 858-5687	exon 5/5 654-5483

Notes:

Exon 1 of mRNA variant (1) and exon 1 of mRNA variant (6) have overlapping 5' ends.

Exon 5 of mRNA variant (1) is with truncated 3'end. The 3'ends of all other mRNA variants are similar.

Mapping of the exons

As the exons represent less than 5% of the pre-mRNA, it is of interest to map their position. Table 1 shows the combinations of exons present in each mRNA variant. The position of the exons on the pre-mRNA sequence was determined (see Materials and Methods) and shown in Fig. 1.

The distribution of the exons along the pre-mRNA is remarkably irregular. Two exons - 1/5 in variant (1) and 1/5 in variant (6) - are at the 5'-end of the primary transcript. Downstream there is a huge gap (intron) of 150 kb with no exons present. The last 10 kb at the 3'-end contain the remaining ten exons.

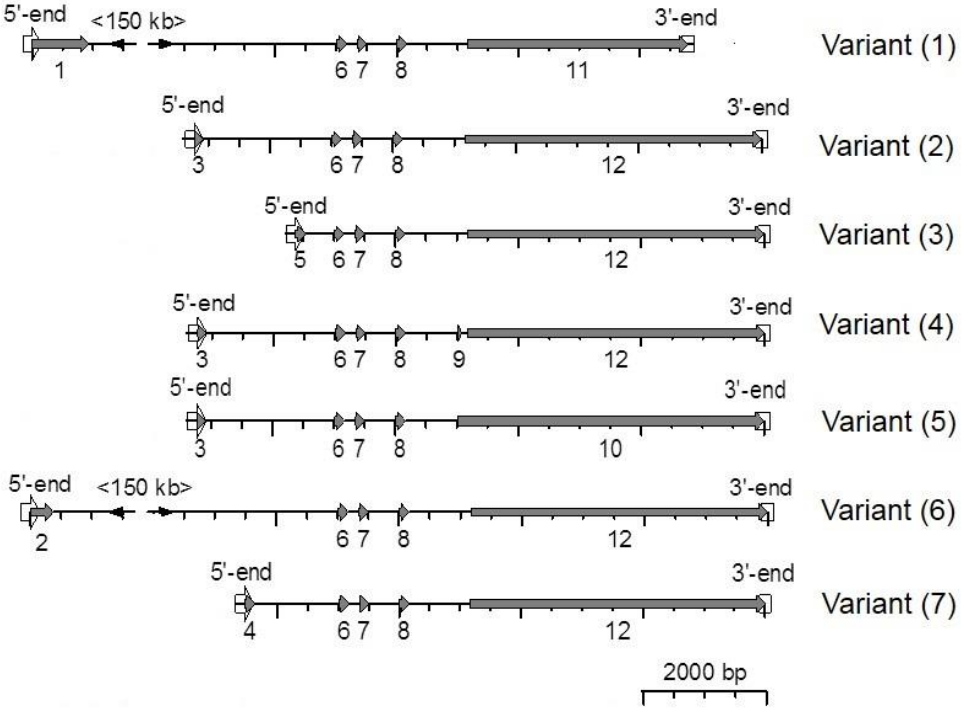

Figure 1. Position of the exons on the primary transcript.

The exons are numbered according to the distance from the start of transcription. The introns are presented as lines, connecting exons. Exons 3 – 12 are shown on the 3'-end of the transcript (12 kb). Exons 1 and 2 are 150 kb upstream.

The 5' untranslated regions

In all eukaryotes, the mRNAs contain 5' and 3' untranslated, non-coding regions. In eukaryotic cells, the translation of the mRNAs starts with the attachment of a scanning complex to the 5'-end of the messenger. The scanning complex (a 40S ribosomal subunit, carrying Met-tRNA$_i$ and translation initiation factors) migrates in the 5' to 3' direction and "inspects" the mRNA for sites suitable for initiation. When an AUG triplet in a proper nucleotide context is encountered, the 60S ribosomal subunit is recruited and the translation is initiated (first-AUG rule).

The introduction of the scanning model (Kozak 1978) increased the interest in the mRNA region upstream of the main open reading frame (the 5'-end untranslated region). Studies showed that in a large fraction of proteins, this region may possess structural features, which may interfere with the scanning process and reduce the efficiency of translation (Kozak 2002).

The examination of the NCBI records showed that translation in all mRNA variants starts at an AUG codon, positioned 15 bp downstream of the 5'-end of the second exon (2/5 or 2/6) (Table 1). Thus, the first exon in all variants is untranslated. It is relevant to note that in the HMGB1 mRNA variants the first exons are different (Table 1 and Fig. 1). This may reflect differences in translation regulation and a method to ensure expression in a variety of tissues and conditions.

The coding exons and termination.

The second exon and the next two exons in the mRNA variants are translated. Accordingly, these three exons are shared by all mRNAs (see Table 1 and Fig. 2). They are small in size, with a combined length of 485 bp, coding for 157 amino acids.

In variants (1) – (3), (6), and (7), translation continues into the next exon (5/5). After 58 codons, a stop codon terminates the open reading frame. Thus, the HMGB1 protein is synthesized as a single polypeptide chain of 215 amino acid residues, coded by 645 nucleotides.

However, in mRNA variants (4) and (5) translation encounters sequences, which in other variants are skipped during splicing. In variant (4), these sequences represent a short exon (50 bp). A phenylalanine residue is added

to the polypeptide chain, then the translation is terminated. In variant (5) the short 50 bp intron is appended to the downstream sequences (Fig. 2).

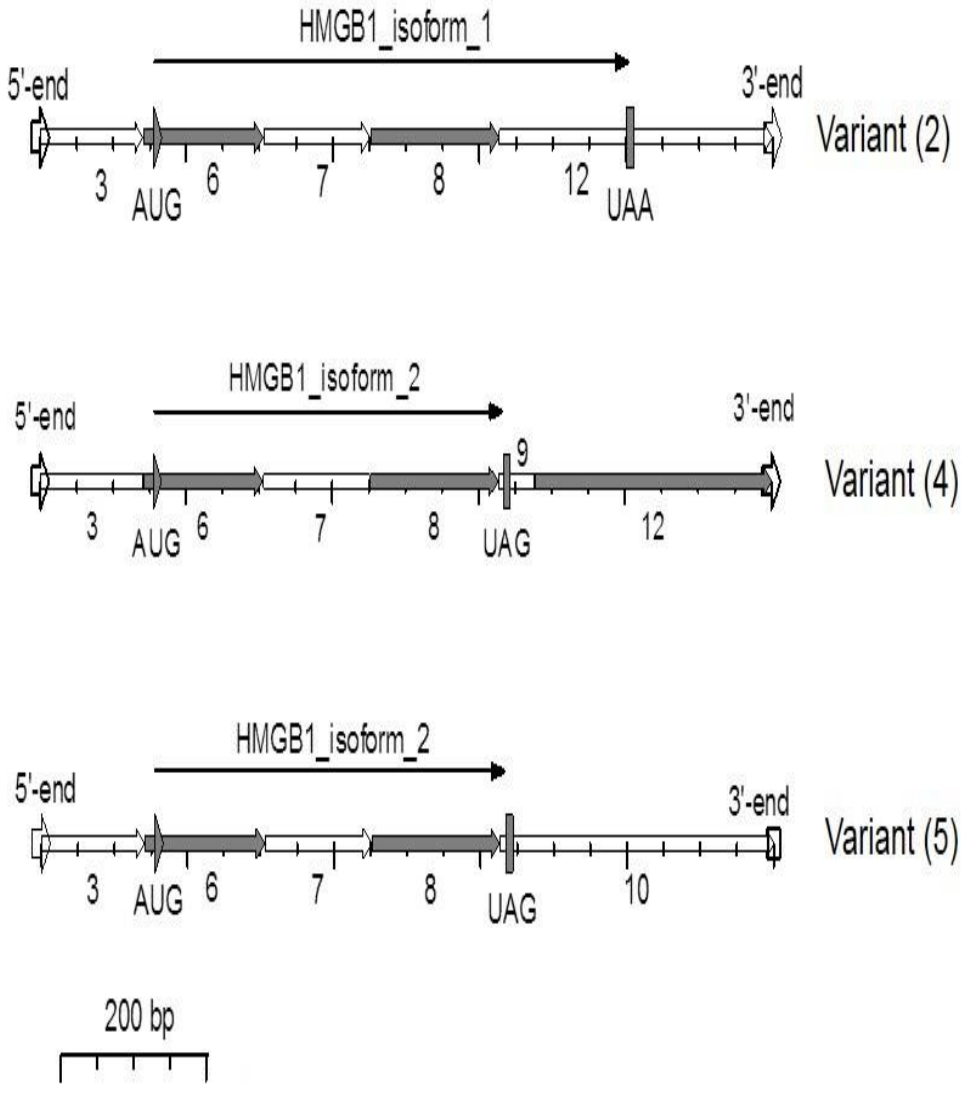

Figure 2. Translation of mRNA variants expressing HMGB1 proteins isoform 1 and isoform 2.

AUG – initiation codon; UAA, UAG – stop codons. The exons composing the messengers are indicated and numbered.

The 3'-end untranslated region

In all HMGB1 variants the 3'-end untranslated region is much longer than the coding part of the messenger. As an example, in variant (3) the distance from AUG to UAA is 647 bp, but the distance from the stop codon (UAA) to the 3'-end is 4653 bp. This is not unusual. Many protein messengers contain long untranslated 3'-end regions, which presumably are targets of regulatory molecules.

The HMGB1 isoforms

The high-mobility group box 1 protein contains 215 amino acid residues, organized in two DNA binding regions and a C-end acidic tail, a flexible structure that can reversibly interact with the HMG boxes and modulate binding to DNA or proteins (see Fig.3). However, the computation analysis showed that two of the transcript variants encode a shorter version of the protein (isoform 2) with 158 amino acids. Isoform 2 is without the acidic tail, which presumably affects the interaction of this protein with other molecules.

This isoform was expressed in transfected cells to study inflammation (Meng et al. 2020).

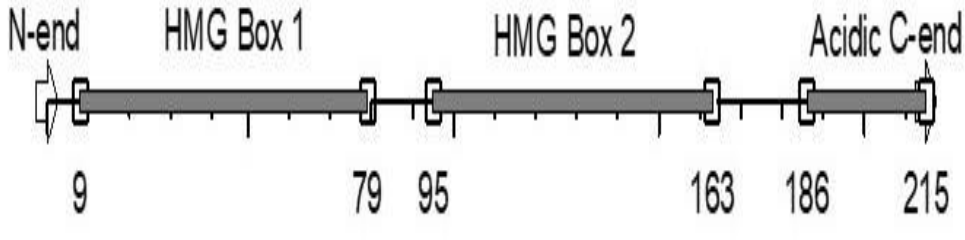

Figure 3. Regions in the HMGB1 protein, isoform 1.

Box HMG1 and Box HMG2 – DNA binding regions. Acidic C-end tail – region composed of aspartate and glutamate residues. This region is absent in HMGB1, isoform 2. The amino acid residues are numbered.

Conclusions

The computation analysis clearly showed the existence of two HMGB1 isoforms. These isoforms are recorded also in the NCBI database of refer-

ence sequences. Never-the-less, the comprehensive protein databanks list only the longer isoform 1. The reasons for that discrepancy are obscure. Some simple experiments are necessary to be designed to clear up this contradiction.

NOTES

1. https://omim.org/entry/163905
2. https://www.ncbi.nlm.nih.gov/gene
3. https://www.uniprot.org/
4. https://www.scied.com

REFERENCES

Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A., 2019. OMIM. org: leveraging knowledge across phenotype–gene relationships. *Nucleic acids research*, **47**(D1), D1038-D1043.

Bianchi, M.E., L. Falciola, S. Ferrari & D.M. Lilley, 1992. The DNA binding site of HMG1 protein is composed of two similar segments (HMG boxes), both of which have counterparts in other eukaryotic regulatory proteins. *EMBO J* 11, 1055-1063.

Bianchi, M.E. & Agresti A ,2005. HMG proteins: dynamic players in gene regulation and differentiation. Current Opinion in Genetics & Development, **15***(5), 496–506.*

Jantzen, H. M., Admon, A., Bell, S. P. & Tjian, R., 1990. Nucleolar transcription factorhUBF contains a DNA-binding motif with homology to HMG proteins. *Nature*, **344**(6269), 830–836.

Lodish, H., Berk, A., Kaiser, C. A., Krieger, M., Scott, M. P., Bretscher, A., Ploegh, H. & Matsudaira, P., 2008. *Molecular cell biology*. Macmillan.

Ferrari, S., Finelli, P., Rocchi, M., & Bianchi, M. E., 1996. The active gene that encodes human high mobility group 1 protein (HMG1) contains introns and maps to chromosome 13. *Genomics*, **35**(2), 367-371.

Kozak, M., 1978. How do eucaryotic ribosomes select initiation regions in messenger RNA? *Cell*, **15**(4), 1109-1123.

Kozak, M., 2002. Pushing the limits of the scanning mechanism for initiation of translation. *Gene*, **299**(1-2), 1-34.

Meng, Y., Qiu, S., Sun, L., & Zuo, J., 2020. Knockdown of exosome-mediated lnc-PVT1 alleviates lipopolysaccharide-induced osteoarthri-

tis progression by mediating the HMGB1/TLR4/NF-κB pathway via miR-93-5p. *Molecular medicine reports*, **22**(6), 5313-5325.

Ugrinova I. & E Pasheva, 2017. HMGB1 protein: a therapeutic target inside and outside the cell. *Advances in protein chemistry and structural biology*, (107), 37-76.

✉ **Lachezar Karagyozov (corresponding author)**
Biological Faculty
Sofia University "St. Kliment Ohridski"
Sofia 1164, Bulgaria
E-mail: lucho3k@gmail.com

✉ **Jordana Todorova**
"Roumen Tsanev" Institute of Molecular Biology
Bulgarian Academy of Sciences
Sofia 1113, Bulgaria
E-mail: jordanabg@yahoo.com