Structural studies of proteins involved in Spinal Muscular Atrophy (SMA): Sam68 and its interaction with RNA

Esther Serrano Pertierra
Máster Universitario en Bioinformática y Bioestadística
Biología Molecular y Estructural

Marta Nadal Rovira
David Merino Arranz

Junio 2018

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Structural studies of proteins involved in Spinal Muscular Atrophy (SMA): Sam68 and its interaction with RNA*
Nombre del autor:	*Esther Serrano Pertierra*
Nombre del consultor/a:	*Marta Nadal Rovira*
Nombre del PRA:	*David Merino Arranz*
Fecha de entrega:	06/2018
Titulación:	*Máster Universitario en Bioinformática y Bioestadística*
Área del Trabajo Final:	*Biología Molecular y Estructural*
Idioma del trabajo:	*Inglés*
Palabras clave	*Spinal muscular atrophy, Sam68, prediction*

Resumen del Trabajo:

La atrofia muscular espinal es una enfermedad neurodegenerativa autosómica recesiva, causada en la mayor parte de los casos por mutaciones en el gen *SMN1*. Los niveles de expresión del gen homólogo, *SMN2*, no compensan la actividad de la proteína SMN. Algunos estudios han descrito que la inestabilidad del gen *SMN2* se debe a una transición C→T en el exón 7, lo cual produce que la mayoría de los tránscritos de *SMN2* se produzcan por empalme alternativo, excluyendo el exón 7.

Sam68 pertenece a la familia de proteínas STAR (transducción de señales y activación del ARN), la cual regula el proceso de empalme alternativo en varios genes implicados en la neurogénesis. La inhibición de la actividad de Sam68 puede rescatar la actividad de SMN, lo que sugiere que esta proteína juega un papel importante en esta enfermedad. El desarrollo de péptidos pequeños / moléculas interferentes para modular el empalme alternativo del gen *SMN2* puede ser efectivo y mejorar la actividad motora. Este enfoque requiere resolver la estructura 3D de Sam68, así como el modelado de la interacción ARN-proteína.

En este proyecto se han llevado a cabo estudios estructurales mediante el uso de diversas herramientas computacionales para la predicción de estructura de Sam68, así como para la predicción de interacciones ARN-proteína. Las puntuaciones obtenidas fueron más altas en la región correspondiente al dominio KH (dominio de unión al ARN), aunque la fiabilidad de la predicción de los residuos que interaccionan con el ARN es más limitada debido a la falta de estudios experimentales disponibles en las bases de datos.

Abstract:

Spinal muscular atrophy (SMA) is an autosomal recessive, and neurodegenerative disease, most of the cases caused by mutations in the *SMN1* gene. The expression levels of the homologue gene, *SMN2*, do not compensate the activity of the SMN protein. A number of studies have described that the instability of the *SMN2* gene is due to a C→T transition in exon 7, which causes that the majority of *SMN2* transcripts are alternatively spliced, excluding exon 7.

Sam68 is a member of the STAR family of proteins (signal transduction and activation of RNA), which regulates the alternative splicing of several genes involved in the neurogenesis. Inhibition of Sam68's activity can rescue SMN activity, suggesting an important role of this protein in the disease. The development of short interfering peptides / molecules to modulate the alternative splicing of the *SMN2* gene may be effective and improve the motor function. This approach requires the resolution of the 3D-structure of Sam68, as well as the modeling of the protein-RNA interaction.

In this project, structural studies Sam68 have been carried out by using different computational tools for protein structure prediction, as well as for prediction of RNA-protein interactions. The scores obtained were greater in the area corresponding to the KH domain (RNA-binding domain), although the reliability prediction of RNA-interacting residues was more limited due to the lack of experimental research available in the databases.

Table of contents

1. Introduction 1
  1.1 Background and motivation for the proposed work 1
    1.1.1. Spinal Muscular Atrophy 1
    1.1.2. Sam68: a member of the STAR family 3
    1.1.3. Sam68 – RNA interactions 5
    1.1.4. Current treatments of spinal muscular atrophy 6
    1.1.5. Protein structure and modeling in drug discovery and design 8
  1.2. Objectives 9
  1.3 Materials and methods 9
    1.3.1. Databases 9
    1.3.2. Sequence alignments 10
    1.3.3. Secondary structure prediction 10
    1.3.4. Model prediction 10
    1.3.5. RNA-protein interactions 12
    1.3.6. Software 12
  1.4. Project scheduling 13
  1.5. Summary of the results 14
  1.6. Summary of the chapters in the Masters Dissertation 15
2. Structure of the protein Sam68 16
  2.1. Primary structure 16
  2.2. Secondary structure 16
    2.2.1. Prediction of the secondary structure 17
  2.3. Experimental structures and models 20
    2.3.1. 3QHE 21
    2.3.2. 2XA6 22
    2.3.3. Structures provided by Feracci *et al*. 23
    2.3.4. Summary of the study of PDB structures 31
  2.4. Domains of Sam68 34
3. Structure predictions of Sam68 35
  3.1. Preliminary aspects 35
  3.2. Homology based predictions 36
    3.2.1. HHPRED 36
    3.2.2. BhageerathH 37
    3.2.3. RAPTOR X 40
    3.2.4. Phyre 2 41
    3.2.5. IntFOLD3 42
    3.2.6. SWISS-MODEL 43
    3.2.7. Summary of the predicted models of Sam68 43
  3.3. Protein threading 46
    3.3.1. I-TASSER 46
  3.4. *Ab initio* predictions 47
    3.4.1. QUARK 47
  3.5. Summary 48
4. Sam68 – RNA interactions 52
  4.1. Predictions of RNA-binding residues from protein sequence 53
  4.2. Predictions of RNA-binding residues from protein and RNA sequences 54

4.3. Prediction of RNA-binding sites from protein structures 55
4.4. Limitations of the study of RNA – protein interactions 59
5. Conclusions 60
6. Glossary 62
7. References 63
8. Appendices 70
8.1. PDB structures used as templates for homology modeling 70
8.1.1. 4JVH 70
8.1.2. 4JVY 72
8.1.3. 2BL5 74
8.1.4. 4WAL 76
8.2. SMN2 transcripts 78
8.3. Amino acid nomenclature 80

List of figures

Figure 1. 1
Figure 2. 2
Figure 3. 3
Figure 4. 7
Figure 5. 8
Figure 6. 16
Figure 7. 17
Figure 8. 18
Figure 9. 20
Figure 10. 22
Figure 11.. 23
Figure 12. 23
Figure 13.. 24
Figure 14. 25
Figure 15. 26
Figure 16.. 27
Figure 17. 28
Figure 18. 29
Figure 19.. 30
Figure 20. 31
Figure 21. 32
Figure 22. 33
Figure 23. 34
Figure 24.. 35
Figure 25.. 36
Figure 26. 37
Figure 27. 37
Figure 28. 38
Figure 29. 38
Figure 30. 39
Figure 31. 40
Figure 32. 41
Figure 33. 41
Figure 34. 42
Figure 35. 43
Figure 36. 44
Figure 37. 46
Figure 38. 48
Figure 39. 49
Figure 40. 51
Figure 41. 52
Figure 42. 53
Figure 43. 54
Figure 44. 56
Figure 45. 57

Figure 46.. 58
Figure 47. 59
Figure 48. 59
Figure 49. 59
Figure S1. 70
Figure S2. 71
Figure S3. 72
Figure S4. 73
Figure S5. 74
Figure S6. 75
Figure S7. 76
Figure S8. 77

List of tables

Table 1.. 2
Table 2.. 18
Table 3.. 20
Table 4.. 44
Table 5.. 45
Table 6.. 47
Table 7.. 49
Table 8.. 53

1. Introduction

1.1 Background and motivation for the proposed work

1.1.1. Spinal Muscular Atrophy

Spinal Muscular Atrophy (SMA) is an autosomal recessive, and neurodegenerative disease. It is characterized by degeneration of the α-motoneurons of the anterior horn in the spinal cord (Figure 1), which leads to progressive weakness and muscular atrophy; in the most severe cases, it could cause paralysis. Currently, it is divided into three main subtypes, being the Type I the most severe. There is a fourth subtype, Type 0, with onset in the gestational period and death before six months of age. On the contrary, Type IV is the mildest form with adult-onset and an average life expectancy [1]. A summary of the clinical classification of SMA is shown in Table 1.

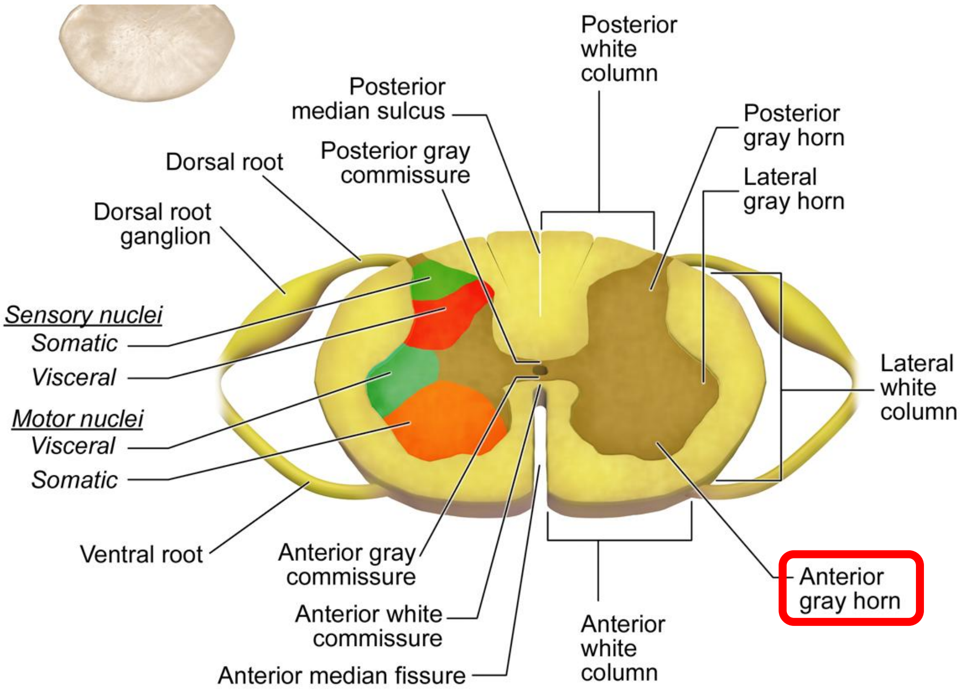


Figure 1. Cross section of the spinal cord. The SMA affects α-motoneurons of the anterior horn.

Up to date there are only a small number of epidemiological studies, although it is commonly cited an incidence of about 1 in 10000 births. For more information about epidemiological data of SMA, see the work of Verhaart *et al.* [2].

SMA Type	Natural age of death
0	< 6 months
I **(Werdnig Hoffmann disease)**	< 2 years
II	> 2 years
III **(Kugelberg Welander disease)**	Adulthood
IV	Adult

Table 1. Clinical types of Spinal Muscular Atrophy.

SMA is an autosomal recessive disease, most of the cases caused by mutations in the *SMN1* gene, located in the telomeric region of the chromosome 5q13.2 [3]. The *SMN1* gene encodes the SMN (survival motoneuron) protein, which regulates the byosinthesis of the small nuclear riboproteins. A homologue gene, *SMN2*, is present in humans. However, the expression levels of this gene do not compensate the activity of the SMN protein. A number of studies have described that the instability of the *SMN2* gene is due to a C→T transition in exon 7, which causes that the majority of *SMN2* transcripts are alternatively spliced, excluding exon 7. As a result, the protein has a different C-terminal end, unstable and non-functional (Figure 2), and does not support the survival of α-motoneurons [3, 4].

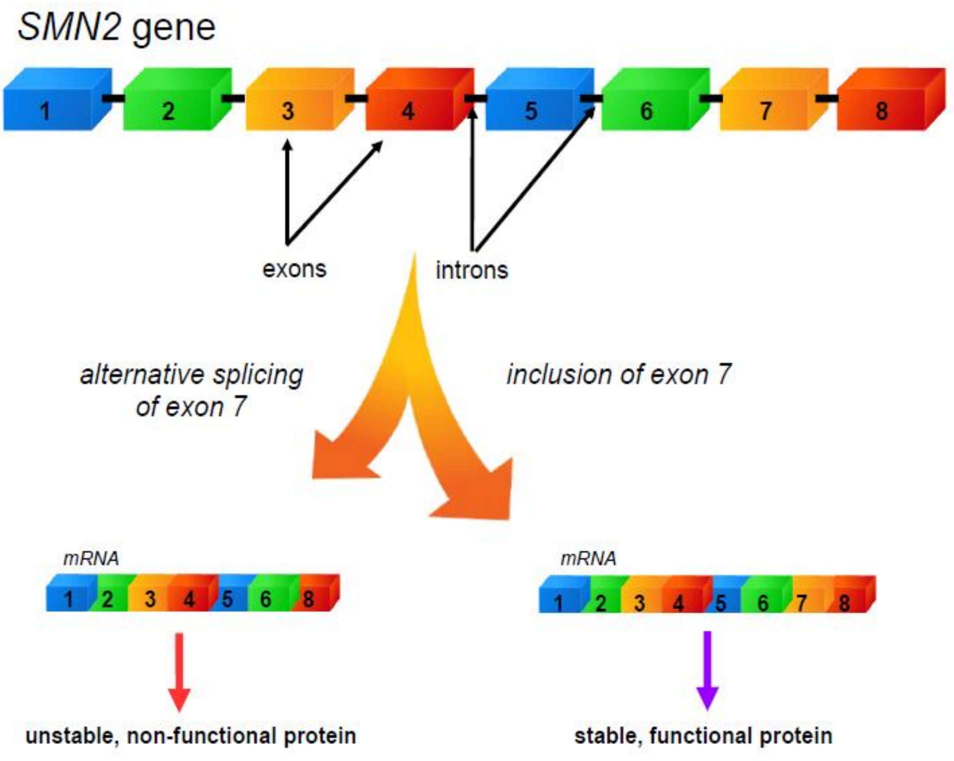

SMN2 gene
1
2
3
4
5
6
7
8
exons
introns
alternative splicing of exon 7
inclusion of exon 7
mRNA
1 2 3 4 5 6 8
mRNA
1 2 3 4 5 6 7 8
unstable, non-functional protein
stable, functional protein

Figure 2. Schematic representation of the alternative splicing of exon 7 in the *SMN2* gene.

1.1.2. Sam68: a member of the STAR family

RNA-binding proteins (RBP) generally regulate alternative splicing. The work of Lukong and Richard [5] showed a potential binding site of the RBP Sam68 (Src-associated protein in mitosis of 68 kDa) by sequencing of exon 7, located upstream the consensus sequence for the heterogeneous nuclear ribonucleoprotein hnRNP A1. Sam68 (SRC associated in mitosis of 68 kDa) is a member of the STAR family (signal transduction and activation of RNA metabolism), which is a family of RNA binding proteins (RBP) that link signaling pathways to DNA processing [5], including the alternative splicing of several genes involved in the neurogenesis.

1.1.2.1. Structure of Sam68

Proteins belonging to the STAR family share common structures. They contain the GRP33/SAM68/GLD-1 (GSG) domain for RNA binding, intracellular localization, and homodimerization domain. It is formed by a heterogeneous nuclear ribonucleoprotein particle K (hnRNP K) homology domain (KH), which is flanked by two regulatory regions. The KH RNA binding domain consists of 70-100 amino acids and it is a region evolutionarily preserved [6]. Sam68 also contains six proline-rich motifs (P0-P5), arginine/glycine/glycine (RGG) and arginine/glycine (RG) boxes, a tyrosine rich region at the C-terminal domain, and a nuclear localization signal (NLS). Figure 3 shows the structure of Sam68.

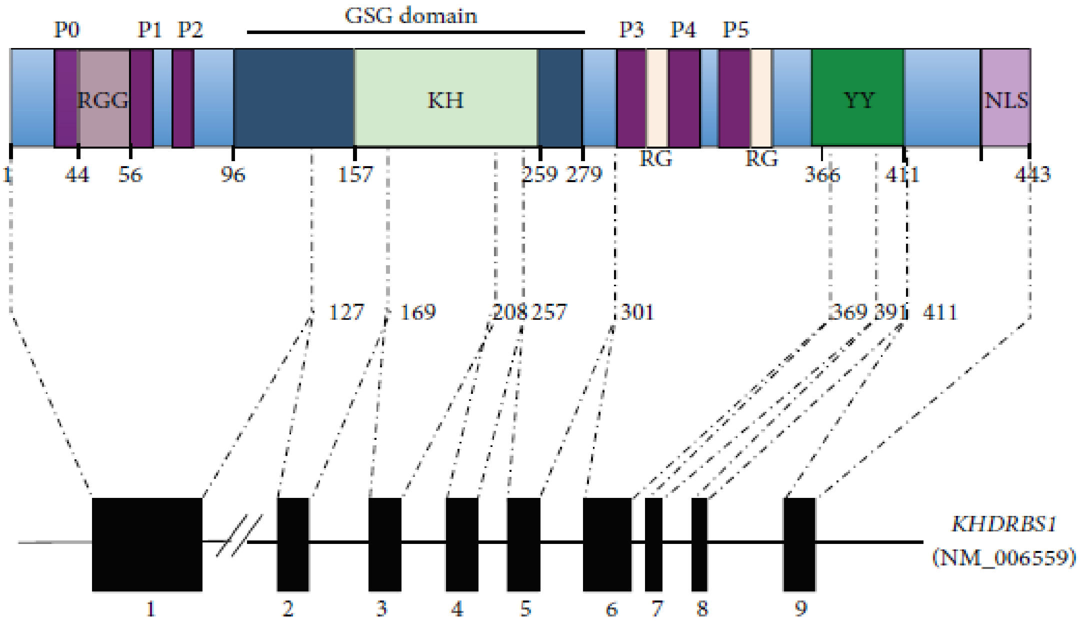


Figure 3. Schematic representation of the protein Sam68 and one of the mRNA isoforms of Sam68 (taken from Frisone *et al.* [7]).

1.1.2.2. Sam68 and apoptosis

The RBP Sam68 has been implicated in the regulation of apoptosis [8, 9]. The work of Paronetto *et al*. [10] described that Bcl-x is a target of Sam68. Alternative splicing results in the production of either the proapoptotic Bcl-x(L) protein or the antiapoptotic Bcl-x(L) protein. They reported that the levels of Sam68 affected the ratio between Bcl-x(L) and Bcl-x(s) proteins, by shifting the balance toward the proapoptotic isoform when intracellular levels of Sam68 increased. Point mutations or knock-down of Sam68 impaired its activity regarding the alternative splicing of Bcl-x. Posttranslational modifications differentially regulate Sam68 activity. In the case of Bcl-x, phosphorylation by Fyn promotes the expression of Bcl-x(L), whereas phosphorylation by Erk1/2 does not affect Sam68-mediated alternative splicing of Bcl-x.

1.1.2.3. Sam68 and cancer

Gene expression and post-transcriptional regulation are processes commonly modified in cancer cells, accompanied by alterations in alternative splicing and translation. Sam68 has been associated with different types of cancer. In prostate cancer patients, higher levels of Sam68 were reported and knock-down of Sam68 by RNAi affected cell proliferation and survival [11]. High Sam68 expression was also found in renal cell carcinoma which, together with its cytoplasmic localization, may be associated with prognosis and survival in this disease [12]. Similar findings regarding mRNA levels, protein levels and cytoplasmic localization were reported by Li *et al.* [13] in cervical cancer. Moreover, down-regulation of Sam68 by shRNA impaired motility and invasiveness of cancer cells. Levels of Sam68 were increased at both transcriptional and translational levels in oral tongue cancer and can be related to disease prognosis as well [14]. Upregulation of Sam68 was found in colorectal cancer (CRC) cell lines and CRC patients, suggesting an oncogenic role of this protein [15].

Taken together, these works and others support the hypothesis that Sam68 play an oncogenic role and that it can be used as a prognostic tool in several types of cancer.

1.1.2.4. Sam68 and spinal muscular atrophy

Although it has been described the implication of Sam68 in several types of cancer, up to date there are only a few studies reporting an association of Sam68 and SMA.

Inhibition of Sam68 activity can rescue SMN activity [16], suggesting an important role of this protein in the disease. In addition, several compounds which interfere *SMN2* splicing have been shown to improve the motor function and to increase SMN protein levels in mice [17].

1.1.3. Sam68 – RNA interactions

As mentioned above, members of the STAR family have a highly conserved RNA-binding domain. It consists of a KH domain delimited by the QUA1 and QUA2 regions, also referred as N-terminal of KH (NK) and the C-terminal of KH (CK) regions, respectively [5].

Sam68 binds to poly(U) [18] and poly(A) [6] ribonucleotide homopolymers. Further experiments using SELEX (systematic evolution of ligands by exponential enrichment) approaches identified UAAA, UUUA, and a bipartite UAAA-UUAA motif as targets for Sam68 [19, 20].

More recently, interaction of Sam68 with long non-coding RNA (lncRNA) has also been described [21]. Specifically with a noncoding RNA activated by DNA (NORAD), which binds to RBPs and affects their ability to regulate other targets. This work described that the interaction with Sam68 is required for recruitment of the gen *PUM2* to NORAD, regulation of Pum activity and proper chromosome segregation.

Regarding the involvement of Sam68 in SMA, an interaction of Sam68 and the *SMN2* gene was reported [16]. Specifically, an exonic splicing silencer (ESS) in exon 7, UUUUA, is a binding site for Sam68. Recently, Feracci *et al.* [22] provided a high-resolution structure of Sam68 (and T-STAR protein) bound to its target RNAs and showed that the QUA1 region is involved in Sam68 dimerization. Moreover, the optimal sequence is a (A/U)AA-$N_{>15}$- (A/U/AA) bipartite motif, with a linker length of more than 15 nucleotides between the two parts of the motif. The different structures from this work, in complex with RNA have been deposited in the PDB database with the accession numbers: 5EL3 (T-STAR KH free), 5ELR (T-STAR KH-QUA2/AAUAAU), 5ELS (T-STAR KH/AAAUAA), 5ELT (T-STAR QUA1-KH/UAAU) and 5EMO (T-STAR STAR/AUUAAA). It is noteworthy to mention that these structures are based on T-STAR (also named SML-2), another member of the STAR family which shares around 70% amino acid sequence identity with Sam68.

To date, this is the only work which has addressed the structural study of the protein Sam68 and its interaction with RNA. Since the regulation of the alternative splicing in *SMN2* gene, in which Sam68 is involved, has emerged as a promising target for therapeutic approaches, more structural studies of Sam68 and its interaction with RNA are needed.

1.1.4. Current treatments of spinal muscular atrophy

SMA is characterized for the loss of the *SMN1* gene, encoding for the SMN protein. A duplicate gene *SMN2* does not compensate the functional loss of *SMN1* in SMA patients.

The SMN protein is ubiquitously expressed and it is localized in the cytoplasm or the nucleoplasm (Figure 4). Several functions have been attributed to this protein, including the assembly of small nuclear ribonucleoproteins (snRNP) [23], synaptic vesicle release [24], mRNA transport [25], or cytoskeleton dynamics [26, 27]. Nevertheless, despite the several processes in which SMN is involved, none of them is exclusively responsible of the SMA disease.

Currently, the therapeutic options for SMA can mainly be divided into SMN-dependent or SMN-independent approaches. Due to the aim of this project, we will focus on SMN-dependent approaches, although the extended review in SMA therapeutics of Bowermann *et al.* [28] is strongly recommended.

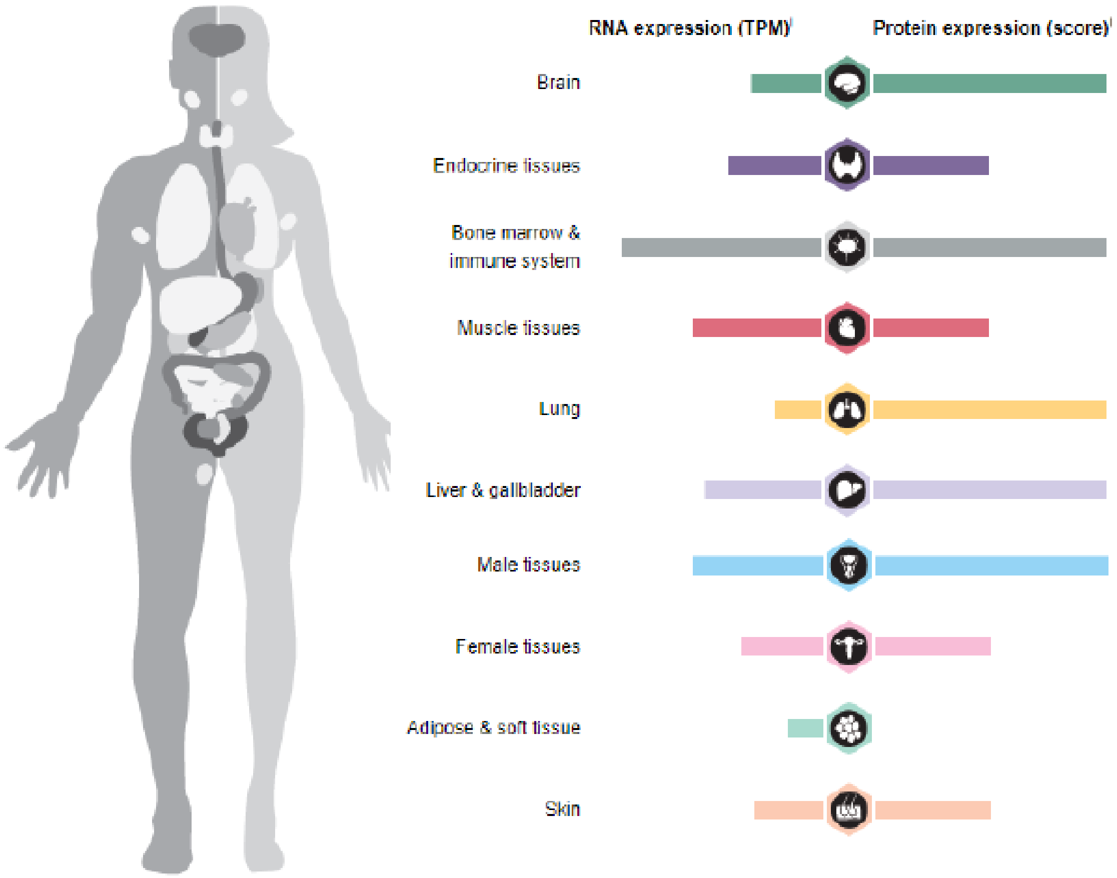


Figure 4. An overview of RNA and protein expression of the *SMN1* gene and the SMN protein, respectively. Data of RNA expression are reported as numbers of transcripts per million (TPM). The protein scores are based on a best estimate of the "true" protein expression from a knowledge-based annotation. This overview can be accessed through https://www.proteinatlas.org.

As previously mentioned, the instability of the *SMN2* gene is due to a C→T transition in exon 7, which eventually causes the alternative splicing of the transcripts, excluding exon 7. Therefore, interfering with this process has become a new therapeutic strategy, allowed by the development of the antisense oligonucleotide (ASO) technology [29].

Singh *et al.* [30] showed that a sequence in intron 7 of the *SMN2* gene (intron splicing silence N1, ISS-N1) may favor the exclusion of the exon 7. This finding has led to the development of an ASO therapy based on this sequence [31] named nusinersen (commercial name Spinraza). The drug was approved in December 2016 by the United States Food and Drug Administration (US FDA) for all types of SMA. Nusinersen was also recommended for European Union approval by the EMA in April 2017 and given marketing authorization in June 2017 [28]. Figure 5 shows the ISS-N1 sequence and the site where nusinersen anneals.

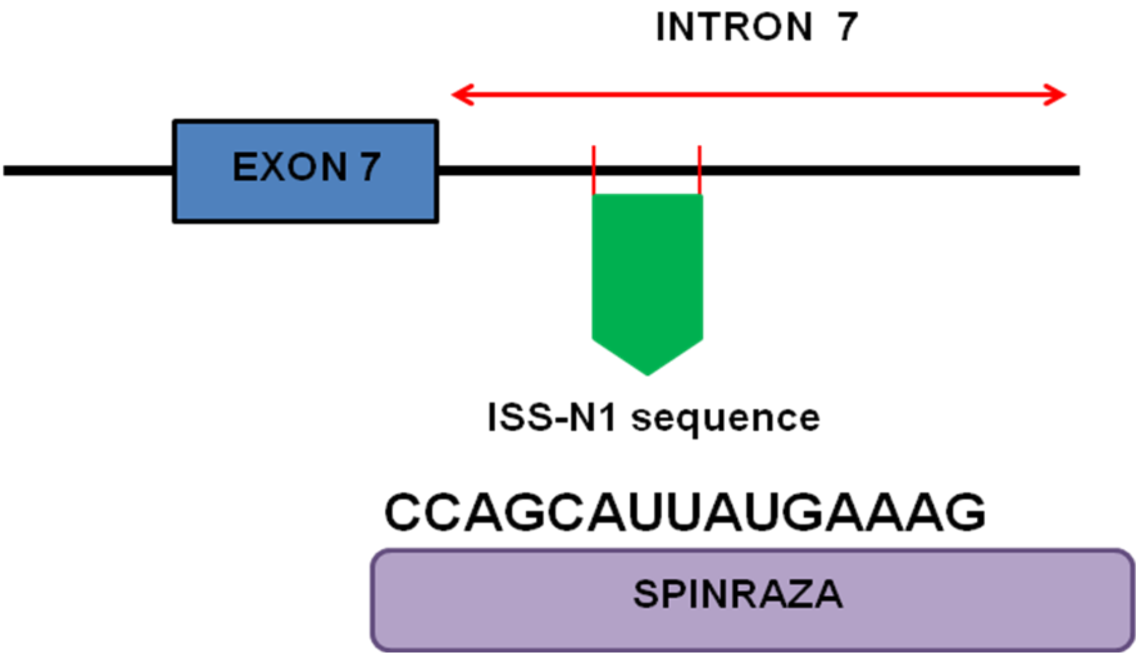


Figure 5. Schematic representation of the region within the ISS-N1 sequence where the the nusinersen anneals to prevent the exclusion of exon 7.

1.1.5. Protein structure and modeling in drug discovery and design

Structural Bioinformatics (or Computational Structural Biology) is the branch of bioinformatics which is related to the analysis and prediction of the three-dimensional structure of biological macromolecules such as proteins, RNA, and DNA.

In the last years, the number of sequences and structure data of proteins has increased exponentially. The improvement in experimental and computational methods has contributed to the development of huge Data Banks where the current knowledge about protein sequence, structure, function, and interaction is collected.

This field of expertise has paved the way for a better understanding of protein interactions with other molecules, and therefore for new methods of drug design.

In the case of SMA, several pieces of evidence point to a key role of Sam68 in SMA and to consider this protein as an emerging therapeutic target. The development of short interfering peptides / molecules to modulate the alternative splicing of the *SMN2* gene may be effective and improve the motor function in SMA patients. This approach requires the resolution of the 3D-structure of Sam68, as well as the modeling of the protein-RNA interactions.

1.2. Objectives

The aims of this project are:

-Understanding of the role of Sam68 in spinal muscular atrophy.

-Studying the Sam68 structure.

-Analysis of currently known experimental structures related to Sam68.

-Using different computational tools to predict Sam68 3D-structure.

-Using different computational tools to predict Sam68-RNA interactions.

1.3 Materials and methods

1.3.1. Databases

1.3.1.1. UniProt

This protein database (http://www.uniprot.org/) provides a large amount of annotations in comparison with any other protein databases. This database was used to obtain information about protein function, subcellular location, sequence, structure, family domains, etc.

1.3.1.2. PDB

The Protein Bank Database (PDB) is the only international repository of protein structures (https://www.rcsb.org/) [32]. It provides information of the experimental data, sequence, ligands, etc. In addition, the structural information, coordinate files, FASTA sequence files, or Fo-fc / 2Fo-fc electron density maps can be downloaded for further analysis. This database was used to collect and study the different PDB structures related to Sam68.

1.3.1.3. PDBsum

Based on the protein structures deposited in the PDB, the PDBsum database (www.ebi.ac.uk/pdbsum) provides images of the PDB structures, secondary structure analysis, or diagrams of protein-ligand and protein-nucleic acid interactions [33]. This database was used to analyze the residues of the PDB structures related to Sam68 in contact with the RNA.

1.3.1.4. InterPro

This database (https://www.ebi.ac.uk/interpro/) provides an analysis of protein sequences, and classifies them into families, domains, and other important sites [34].

1.3.2. Sequence alignments

BLAST protein (BLASTp) was used to find homologous proteins of Sam68 by local alignments. This tool is provided by the National Center for Biotechnology Information (NCBI) at https://blast.ncbi.nlm.nih.gov/Blast.cgi .

1.3.3. Secondary structure prediction

1.3.3.1. PSIPRED

PSIPRED (Psi-blast based secondary structure prediction) was used to visualize the secondary structure prediction of Sam68. This tool is available at http://bioinf.cs.ucl.ac.uk/psipred/.

1.3.3.2. JPRED

JPRED (http://www.compbio.dundee.ac.uk/jpred/index_up.html) was used for secondary structure prediction as well. It uses the algorithm Jnet for the prediction of the secondary structure and also coiled-coils regions and solvent accessibility using the Lupas method [35].

1.3.4. Model prediction

1.3.4.1. Protein disorder prediction

PrDOS server has been used to study the presence of disordered regions in our query protein [36]. This tool scores the disorder probability of each residue and it is available at the website http://prdos.hgc.jp/cgi-bin/top.cgi

1.3.4.2. Signal peptide prediction

Signal peptides are 20 – 30 amino acids fragments present in the majority of newly translated proteins which are going through the secretory pathway. Transmembrane proteins, proteins with subcellular localization (mainly Golgi apparatus, Endoplasmic Reticulum, or endosomes), and proteins to be secreted are targeted with these signal peptides. The SignalP 4.1. Server was used to identify signal peptide cleavage sites (http://www.cbs.dtu.dk/services/SignalP/) [37].

1.3.4.3. Comparative protein modeling

-Homology modeling

It is based on the assumption that homologous proteins have similar structures. The search of templates was carried out using HHPred [38, 39], which finds homologous proteins (https://toolkit.tuebingen.mpg.de/#/tools/hhpred). Below is a list of the servers used for 3D prediction of Sam68:

-BhageerathH - http://www.scfbio-iitd.res.in/bhageerathH+/

-Raptor X [40] - http://raptorx.uchicago.edu/

-Phyre2 [41] - www.sbg.bio.ic.ac.uk/~phyre/

-IntFOLD3 [42] - http://www.reading.ac.uk/bioinf/IntFOLD/

-SWISS-MODEL [43] - https://swissmodel.expasy.org/

-Protein threading

It analyzes the primary structure of a protein and relates it to a 3D structure based on a database of solved structures. It is also called 3D-1D fold recognition.

-I-TASSER

I-TASSER (Iterative Threading ASSEmbly Refinement) [44-46] identifies structural models from PDB by multiple threading alignments and constructs atomic models by iterative template fragment assembly simulations (https://zhanglab.ccmb.med.umich.edu/I-TASSER/).

1.3.4.4. *Ab initio* structure prediction

-QUARK

QUARK builds models by Montecarlo simulations using small fragments (1-20 residues), and it is suitable for proteins without homologous templates in PDB [47, 48].

QUARK server leads the rank of free modeling in CASP9 and CASP10 experiments, as stated in its website (https://zhanglab.ccmb.med.umich.edu/QUARK/).

1.3.5. RNA-protein interactions

- PPRint (Prediction of Protein Interaction) was used for the prediction of residues within the Sam68 sequence which may interact with RNA [49]. This site can be accessed at http://webs.iiitd.edu.in/raghava/pprint/index.html.

-RNABindRPlus is a web application which also predicts the RNA-interacting residues (http://ailab1.ist.psu.edu/RNABindRPlus/).

-catRAPID is a server developed by Bellucci *et al*. [50] which estimates the propensity of RNA-protein interactions. More catRAPID modules have been developed since its creation to accomplish different approaches (http://s.tartaglialab.com/page/catrapid_group).

-The KYG method (http://cib.cf.ocha.ac.jp/KYG/) was used for prediction of RNA-binding sites from protein 3D structures [51].

1.3.6. Software

1.3.6.1. Win*Coot*

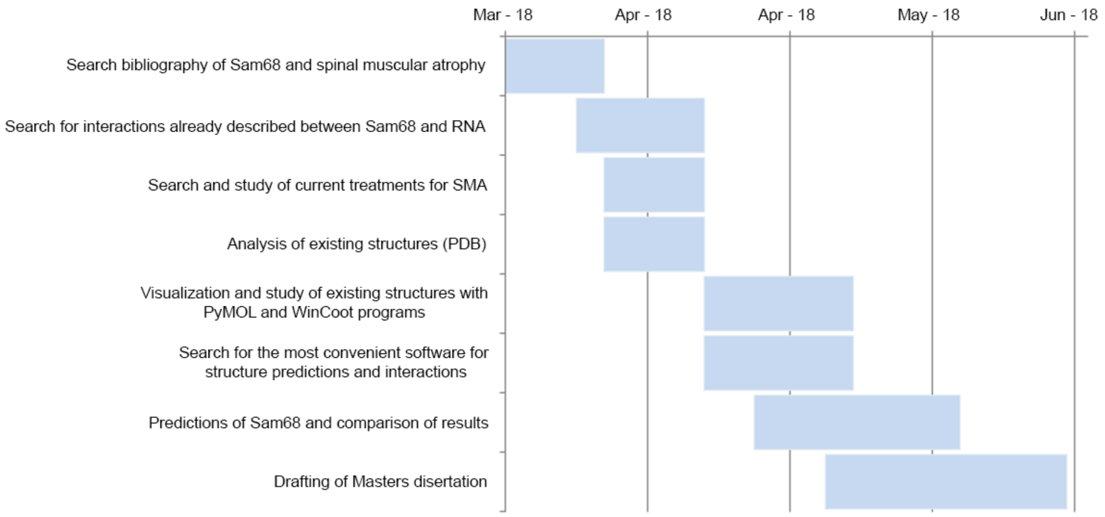
This program is used to display atomic models of macromolecules, enable their manipulation and validation of these models with the 3D electron density maps obtained by X-ray crystallography methods. Win*Coot* was used to visualize the different PDB structures and their electron density maps related to the protein Sam68.

1.3.6.2. PyMOL

PyMOL is one of the most used computer software to visualize from small molecules to macromolecules, and one of the most useful tools in structural biology. It was used to visualize the different PDB structures, as well as the different models generated by the servers used for 3D structure prediction.

1.4. Project scheduling

The main tasks required and their timing to accomplish the aims of this project are summarized in the Gantt diagram shown below:



For each PEC, the specific tasks are detailed according to the initial working plan proposed.

-PEC 2

- Literature review to present and analyse part of the published literature which is relevant for the project: spinal muscular atrophy, Sam68 and its involvement in the disease, and current treatments.
- Search solved protein structures submitted in the Protein Data Bank which are related to Sam68.
- Visualization of these structures using two programs: PyMOL and Win*Coot*.

As a result, the Introduction section was written with the most significant information found in order to provide a proper background and present the aims of this study.

All the current structural information about Sam68 was also collected together with an analysis of the PDB structures available. The program Win*Coot* was used to study these files together with their electron density maps, focusing on those residues in contact with RNA which had lowest scores in the density analysis carried out by the program. PyMOL was used to visualize these structures. As a result, the main section "Structure of the protein Sam68" was drafted and submitted.

-PEC 3

- Search for the most convenient software for structure predictions and interactions.
- Predictions of Sam68 and comparison of the results

The rest of the main chapters of this work derived from these tasks. Different servers were used for the structural prediction of Sam68, in order to compare the results. Homology based predictions and protein threading were the main methods employed for this particular case, since the templates found for the protein had good identity percentages. Nevertheless, an online facility for *ab initio* prediction was also used to check the differences with the previous ones. A similar approach was employed to study RNA-protein interactions.

Once these tasks were completed, the "Materials and methods" section was drafted with all the resources employed to fulfill the purpose of this work. In addition, the "Appendices" section was added with supplemental information and submitted together with the other chapters mentioned above.

-During the Masters Dissertation semester

All the figures and tables required for this work were made in their respective sections. At the end of PEC 2, the drafting of the dissertation and its layout started, including the revision of the "References" and "Glossary" sections.

1.5. Summary of the results

Several predictions of the 3D structure of Sam68 were obtained and visualized, with high reliability results regarding the RNA-binding domain. The areas outside this domain were predicted as highly disordered, and a proper prediction was not achieved.

Interactions between RNA and Sam68 were predicted. However, the accuracy of these predictions is not high enough yet.

1.6. Summary of the chapters in the Masters Dissertation

Chapter 2 – Structure of the protein Sam68: with the updated information about the structure of Sam68 and a review of the experimental structures available related to the protein.

Chapter 3 – Structure predictions of Sam68: description of the different models predicted by different methods (homology modeling, protein threading, *ab initio* prediction) and discussion of the results.

Chapter 4 – Sam68 – RNA interactions: with the different predictions of RNA-interacting residues using different tools and discussion of the results.

Chapter 5 – Conclusions: with the highlights of the work and with certain approaches for future work pointed out.

Chapter 6 – Glossary: a full list of the main terms related to the field of expertise, in alphabetical order.

Chapter 7 – References: the full list of works and websites consulted for this work.

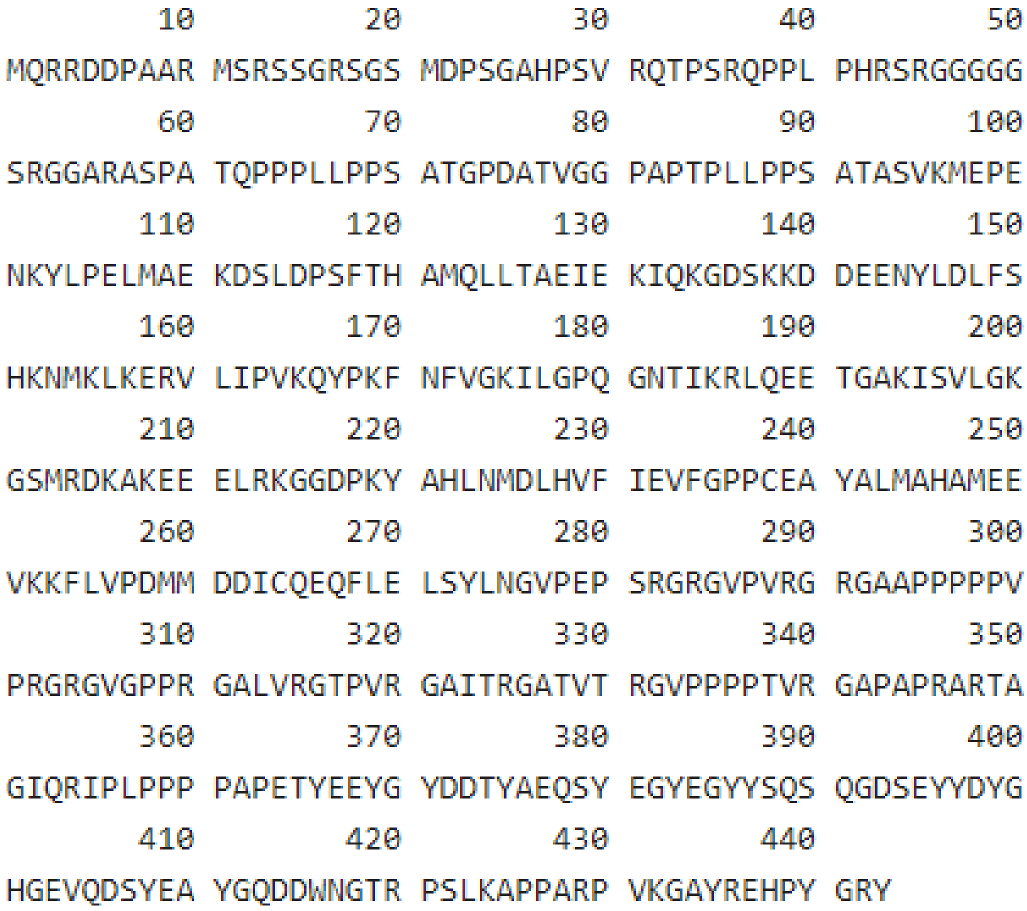
Chapter 8 – Appendices: containing PDB structures used as templates for homology modeling in Chapter 3, but not included in Chapter 2. It also contains the four transcripts know for the *SMN2* gene, and the nomenclature of the amino acids.

2. Structure of the protein Sam68

2.1. Primary structure

The data about the current knowledge regarding the structure of Sam68 have been consulted in UniProt.

Sam68 is a protein of 443 amino acids and *ca*. 48 KDa. The sequence is shown in Figure 6.

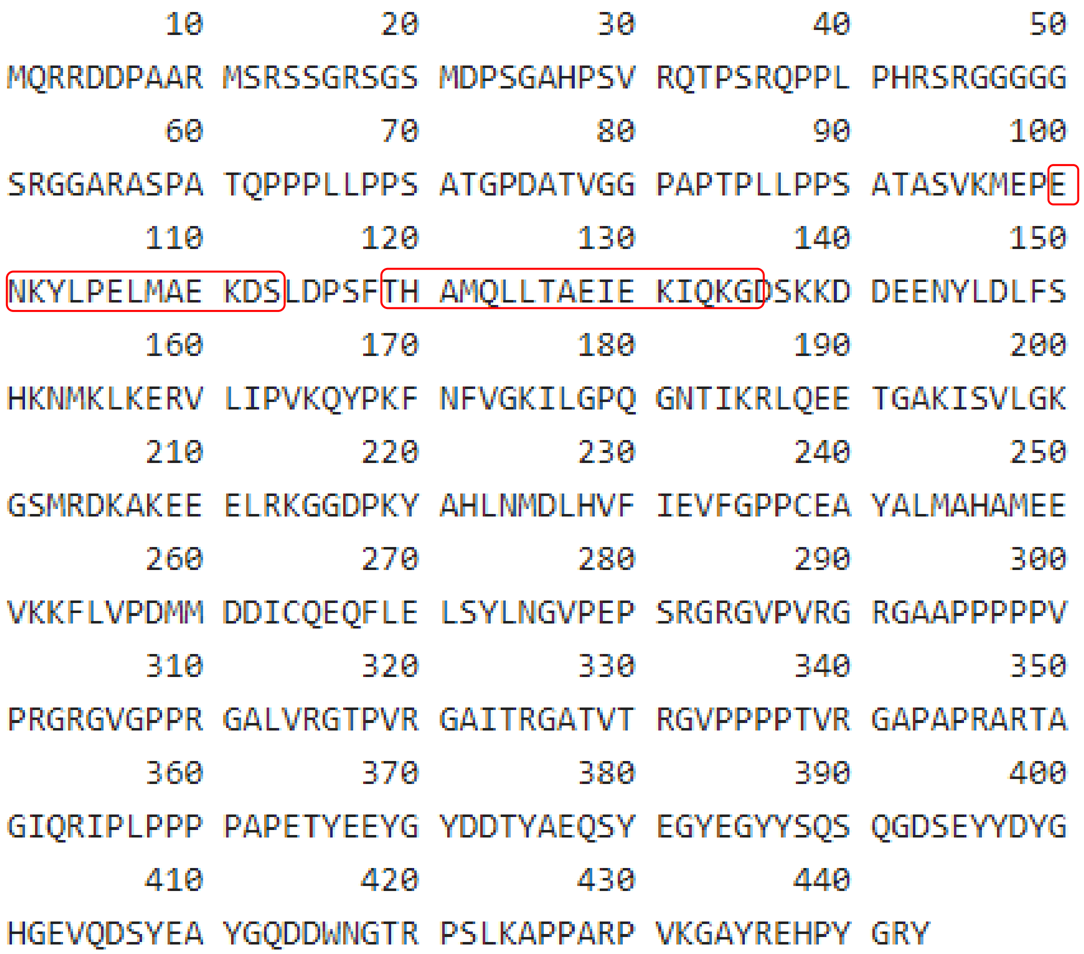


```
        10         20         30         40         50
MQRRDDPAAR MSRSSGRSGS MDPSGAHPSV RQTPSRQPPL PHRSRGGGGG
        60         70         80         90        100
SRGGARASPA TQPPPLLPPS ATGPDATVGG PAPTPLLPPS ATASVKMEPE
       110        120        130        140        150
NKYLPELMAE KDSLDPSFTH AMQLLTAEIE KIQKGDSKKD DEENYLDLFS
       160        170        180        190        200
HKNMKLKERV LIPVKQYPKF NFVGKILGPQ GNTIKRLQEE TGAKISVLGK
       210        220        230        240        250
GSMRDKAKEE ELRKGGDPKY AHLNMDLHVF IEVFGPPCEA YALMAHAMEE
       260        270        280        290        300
VKKFLVPDMM DDICQEQFLE LSYLNGVPEP SRGRGVPVRG RGAAPPPPPV
       310        320        330        340        350
PRGRGVGPPR GALVRGTPVR GAITRGATVT RGVPPPPTVR GAPAPRARTA
       360        370        380        390        400
GIQRIPLPPP PAPETYEEYG YDDTYAEQSY EGYEGYYSQS QGDSEYYDYG
       410        420        430        440
HGEVQDSYEA YGQDDWNGTR PSLKAPPARP VKGAYREHPY GRY
```

Figure 6. Primary structure of Sam68.

2.2. Secondary structure

Two helix motifs are reported in UniProt: one between the amino acids 100-113 (14 amino acids length) and a second between 119-135 residues (16 amino acids length). Figure 7 shows the amino acid sequence with the helix motifs indicated.



```
        10         20         30         40         50
MQRRDDPAAR MSRSSGRSGS MDPSGAHPSV RQTPSRQPPL PHRSRGGGGG
        60         70         80         90        100
SRGGARASPA TQPPPLLPPS ATGPDATVGG PAPTPLLPPS ATASVKMEPE
       110        120        130        140        150
NKYLPELMAE KDSLDPSFTH AMQLLTAEIE KIQKGDSKKD DEENYLDLFS
       160        170        180        190        200
HKNMKLKERV LIPVKQYPKF NFVGKILGPQ GNTIKRLQEE TGAKISVLGK
       210        220        230        240        250
GSMRDKAKEE ELRKGGDPKY AHLNMDLHVF IEVFGPPCEA YALMAHAMEE
       260        270        280        290        300
VKKFLVPDMM DDICQEQFLE LSYLNGVPEP SRGRGVPVRG RGAAPPPPPV
       310        320        330        340        350
PRGRGVGPPR GALVRGTPVR GAITRGATVT RGVPPPPTVR GAPAPRARTA
       360        370        380        390        400
GIQRIPLPPP PAPETYEEYG YDDTYAEQSY EGYEGYYSQS QGDSEYYDYG
       410        420        430        440
HGEVQDSYEA YGQDDWNGTR PSLKAPPARP VKGAYREHPY GRY
```

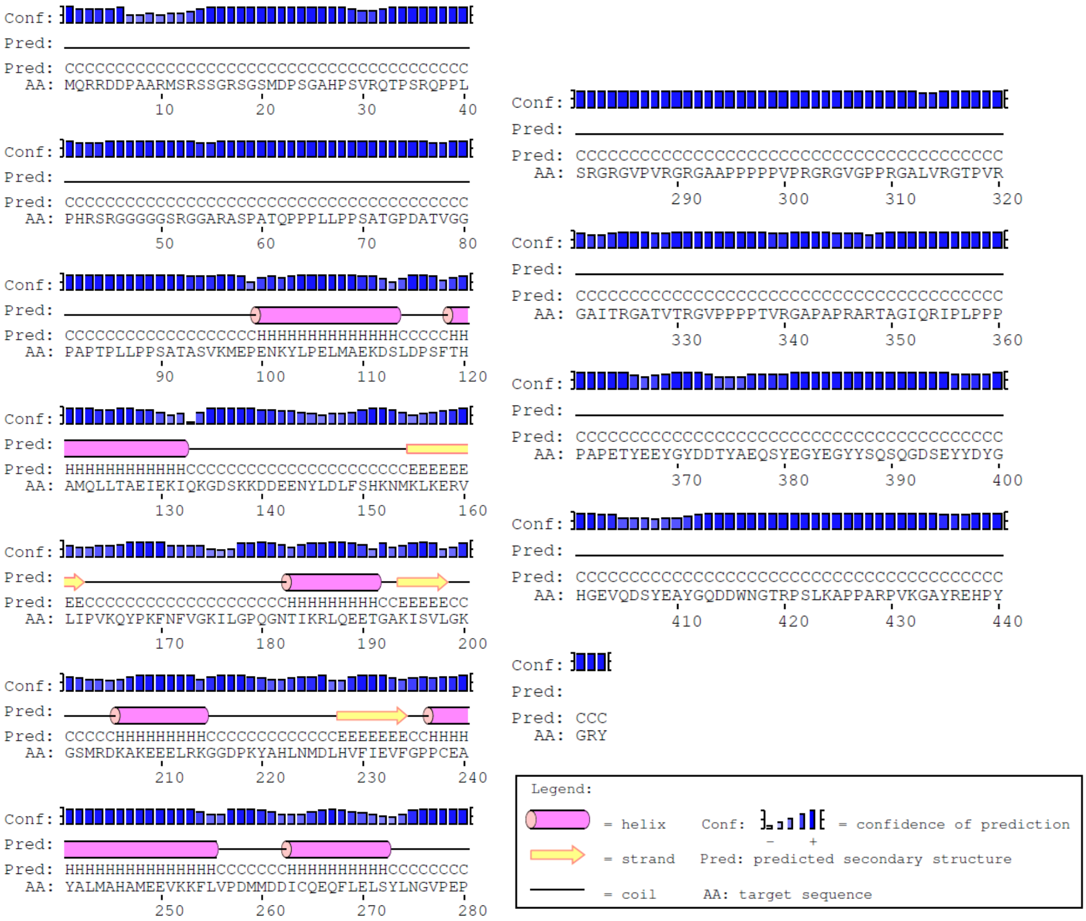
Figure 7. The two helix motifs reported in UniProt are highlighted in red.

2.2.1. Prediction of the secondary structure

2.2.1.1. PSIPRED

PSIPRED (Psi-blast based secondary structure prediction) is one of the best secondary structure predictors, available at http://bioinf.cs.ucl.ac.uk/psipred/.

This tool was used to see a prediction of the secondary structure of Sam68. Figure 8 depicts the output of the prediction. Overall, the confidence of the prediction, indicated by the blue bars, is good. It predicts six α-helix and three β-strands. As it is shown in Figure 8, the α-helix regions indicated in UniProt database are similar to those predicted by PSIPRED, with only one or two residues of difference. Table 2 shows a summary of the structures predicted and where they are located.



```
Conf:
Pred:
Pred: CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
  AA: MQRRDDPAARMSRSSGRSGSMDPSGAHPSVRQTPSRQPPL
               10        20        30        40

Conf:
Pred:
Pred: CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
  AA: PHRSRGGGGGSRGGARASPATQPPPLLPPSATGPDATVGG
               50        60        70        80

Conf:
Pred:
Pred: CCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHCCCCCHH
  AA: PAPTPLLPPSATASVKMEPENKYLPELMAEKDSLDPSFTH
               90       100       110       120

Conf:
Pred:
Pred: HHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCEEEEEE
  AA: AMQLLTAEIEKIQKGDSKKDDEENYLDLFSHKNMKLKERV
              130       140       150       160

Conf:
Pred:
Pred: EECCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHCCEEEEECC
  AA: LIPVKQYPKFNFVGKILGPQGNTIKRLQEETGAKISVLGK
              170       180       190       200

Conf:
Pred:
Pred: CCCCCHHHHHHHHHHCCCCCCCCCCCCCEEEEEEECCHHHH
  AA: GSMRDKAKEEELRKGGDPKYAHLNMDLHVFIEVFGPPCEA
              210       220       230       240

Conf:
Pred:
Pred: HHHHHHHHHHHHHHHHCCCCCCHHHHHHHHHHCCCCCCCC
  AA: YALMAHAMEEVKKFLVPDMMDDICQEQFLELSYLNGVPEP
              250       260       270       280

Conf:
Pred:
Pred: CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
  AA: SRGRGVPVRGRGAAPPPPPVPRGRGVGPPRGALVRGTPVR
              290       300       310       320

Conf:
Pred:
Pred: CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
  AA: GAITRGATVTRGVPPPPTVRGAPAPRARTAGIQRIPLPPP
              330       340       350       360

Conf:
Pred:
Pred: CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
  AA: PAPETYEEYGYDDTYAEQSYEGYEGYYSQSQGDSEYYDYG
              370       380       390       400

Conf:
Pred:
Pred: CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
  AA: HGEVQDSYEAYGQDDWNGTRPSLKAPPARPVKGAYREHPY
              410       420       430       440

Conf:
Pred:
Pred: CCC
  AA: GRY
```

Legend:
= helix
Conf:
- +
= confidence of prediction
= strand
Pred: predicted secondary structure
= coil
AA: target sequence

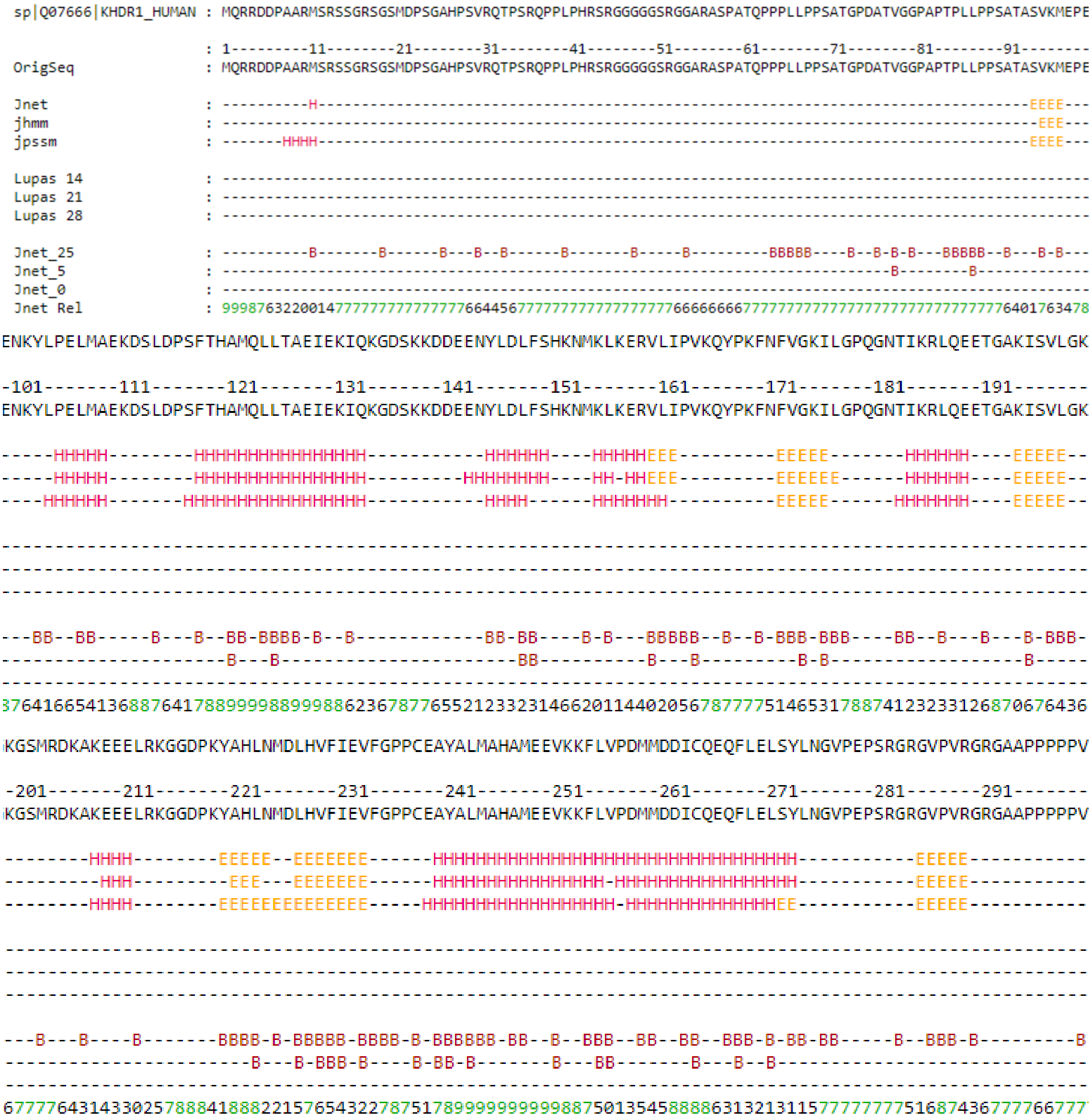
Figure 8. Output of the prediction of the secondary structure of the protein Sam68 using PSIPRED.

α-helix
100-113
119-132
183-191
206-214
237-255
263-273
β-strand
155-162
194-198
228-234

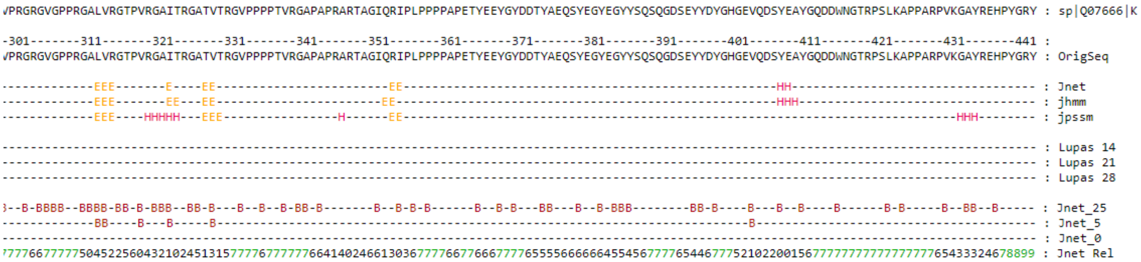
Table 2. Summary of the alpha-helix and beta-strand motifs predicted using PSIPRED and their respective positions in the amino acid sequence.

2.2.1.2. JPRED

JPRED uses the algorithm Jnet for the prediction of the secondary structure and coiled-coils regions and solvent accessibility using the Lupas method [35]. A summary of the results for prediction of Sam68 secondary structure is shown in Figure 9. Some of the structures predicted are identical or similar to PSIPRED prediction, being the α-helix 118-133 and part of the helix in 240-273 the ones with greatest scores. However, JPRED did not predict any coiled-coils regions, as the dashes in Lupas prediction indicate less than 50% probability.



```
sp|Q07666|KHDR1_HUMAN : MQRRDDPAARMSRSSGRSGSMDPSGAHPSVRQTPSRQPPLPHRSRGGGGGSRGGARASPATQPPPLLPPSATGPDATVGGPAPTPLLPPSATASVKMEPE

                      : 1---------11--------21--------31--------41--------51--------61--------71--------81--------91--------
OrigSeq               : MQRRDDPAARMSRSSGRSGSMDPSGAHPSVRQTPSRQPPLPHRSRGGGGGSRGGARASPATQPPPLLPPSATGPDATVGGPAPTPLLPPSATASVKMEPE

Jnet                  : ----------H----------------------------------------------------------------------------------EEEE---
jhmm                  : ----------------------------------------------------------------------------------------------EEE---
jpssm                 : -------HHHH-----------------------------------------------------------------------------------EEEE---

Lupas 14              : ----------------------------------------------------------------------------------------------------
Lupas 21              : ----------------------------------------------------------------------------------------------------
Lupas 28              : ----------------------------------------------------------------------------------------------------

Jnet_25               : ----------B-------B------B---B--B------B-------B-----B---------BBBBB----B--B-B-B---BBBBB--B---B-B---
Jnet_5                : -----------------------------------------------------------------------------B--------B-------------
Jnet_0                : ----------------------------------------------------------------------------------------------------
Jnet Rel              : 9998763220014777777777777777664456777777777777777777777666666667777777777777777777777777776401763478

ENKYLPELMAEKDSLDPSFTHAMQLLTAEIEKIQKGDSKKDDEENYLDLFSHKNMKLKERVLIPVKQYPKFNFVGKILGPQGNTIKRLQEETGAKISVLGK

-101-------111-------121-------131-------141-------151-------161-------171-------181-------191-------
ENKYLPELMAEKDSLDPSFTHAMQLLTAEIEKIQKGDSKKDDEENYLDLFSHKNMKLKERVLIPVKQYPKFNFVGKILGPQGNTIKRLQEETGAKISVLGK

-----HHHHH--------HHHHHHHHHHHHHHHH-----------HHHHHH----HHHHHEEE---------EEEEE-------HHHHHH----EEEEE--
-----HHHHH--------HHHHHHHHHHHHHHHH---------HHHHHHHH----HH-HHEEE---------EEEEEE------HHHHHH----EEEEE--
----HHHHHH-------HHHHHHHHHHHHHHHHH-----------HHHH------HHHHHHH----------EEEEE------HHHHHHH----EEEEE--

-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------

---BB--BB-----B---B--BB-BBBB-B--B------------BB-BB----B-B---BBBBB--B--B-BBB-BBB----BB--B---B---B-BBB-
---------------------B---B----------------------BB----------B---B---------B-B------------------B-----
-----------------------------------------------------------------------------------------------------
37641665413688764178899998899988623678776552123323146620114402056787777514653178874123233126870676436

KGSMRDKAKEEELRKGGDPKYAHLNMDLHVFIEVFGPPCEAYALMAHAMEEVKKFLVPDMMDDICQEQFLELSYLNGVPEPSRGRGVPVRGRGAAPPPPPV

-201-------211-------221-------231-------241-------251-------261-------271-------281-------291-------
KGSMRDKAKEEELRKGGDPKYAHLNMDLHVFIEVFGPPCEAYALMAHAMEEVKKFLVPDMMDDICQEQFLELSYLNGVPEPSRGRGVPVRGRGAAPPPPPV

--------HHHH--------EEEEE--EEEEEEE------HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH-----------EEEEE-----------
---------HHH---------EEE---EEEEEEE------HHHHHHHHHHHHHHHH-HHHHHHHHHHHHHHHHH-----------EEEEE-----------
--------HHHH--------EEEEEEEEEEEEEE-----HHHHHHHHHHHHHHHHHH-HHHHHHHHHHHHHHEE-----------EEEEE-----------

-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------

---B---B----B-------BBBB-B-BBBBB-BBBB-B-BBBBBB-BB--B--BBB--BB--BB--BBB-B-BB-BB-----B--BBB-B---------B
-----------------------B---B-BBB-B----B-BB-B-------B---BB-------B---B--B-----------------------------
-----------------------------------------------------------------------------------------------------
67777643143302578884188822157654322787517899999999998875013545888863132131157777777751687436777766777
```



```
VPRGRGVGPPRGALVRGTPVRGAITRGATVTRGVPPPPTVRGAPAPRARTAGIQRIPLPPPPAPETYEEYGYDDTYAEQSYEGYEGYYSQSQGDSEYYDYGHGEVQDSYEAYGQDDWNGTRPSLKAPPARPVKGAYREHPYGRY : sp|Q07666|K

-301-------311-------321-------331-------341-------351-------361-------371-------381-------391-------401-------411-------421-------431-------441 :
VPRGRGVGPPRGALVRGTPVRGAITRGATVTRGVPPPPTVRGAPAPRARTAGIQRIPLPPPPAPETYEEYGYDDTYAEQSYEGYEGYYSQSQGDSEYYDYGHGEVQDSYEAYGQDDWNGTRPSLKAPPARPVKGAYREHPYGRY : OrigSeq

-------------EEE-------E----EE------------------------EE----------------------------------------------------HH---------------------------------- : Jnet
-------------EEE-------EE---EE-----------------------EE-----------------------------------------------------HHH--------------------------------- : jhmm
-------------EEE----HHHHH---EEE----------------H------EE-----------------------------------------------------------------------------HHH-------- : jpssm

------------------------------------------------------------------------------------------------------------------------------------------------ : Lupas 14
------------------------------------------------------------------------------------------------------------------------------------------------ : Lupas 21
------------------------------------------------------------------------------------------------------------------------------------------------ : Lupas 28

B--B-BBBB--BBBB-BB-B-BBB--BB-B---B--B--B-BB-B-------B--B-B-B------B--B-B---BB---B--B-BBB--------BB-B----B---B--B----B------B-B-----B--BB--B----- : Jnet_25
-------------BB----B---B-----B--------------------------------------------------------------------------B--------------------------------------- : Jnet_5
------------------------------------------------------------------------------------------------------------------------------------------------ : Jnet_0
777766777775045225604321024513157777677777766414024661303677776677666777765555666666455456777765446777521022001567777777777777777765433324678899 : Jnet Rel
```

Figure 9. Output of the secondary prediction using JPRED. Jnet indicates the final secondary structure prediction for the sequence. Lupas indicate de Lupas coil prediction. Below are shown the Jnet predictions of burial of the amino acid (less than 25% solvent accessibility, less than 5% or 0% exposure), and Jnet Rel indicates the reliability of the prediction accuracy, ranging from 0 to 9 (highlighted in green those with good scores).

2.3. Experimental structures and models

The Protein Model Portal (PMP, https://www.proteinmodelportal.org/) was used to access the different models available of this protein by comparative modeling methods. The search of Sam68 in this portal showed two experimental structures and four experimental models covering part of the sequence (Table 3).

MODELS			
Provider	**Templates**	**% Seq. ID**	**From - To**
SWISSMODEL	5ELTA	75 %	100 - 256
MODBASE	5ELTB	74 %	100 - 256
SWISSMODEL	2BL5A	52 %	155 - 282
SWISSMODEL	4JVHA	43 %	98 - 278
EXPERIMENTAL STRUCTURES			
PDB	**Description**	**% Seq. ID**	**From - To**
3QHE	Crystal structure of the complex between the armadillo repeat domain of APC and the tyrosine-rich domain of Sam68	100 %	365 - 419
2XA6	Structural basis for homodimerization of the Src-associated during mitosis 68 kd protein (sam68) QUA1 domain	100 %	97 - 135

Table 3. Models and experimental structures related to Sam68, obtained in the Protein Model Portal.

Besides these structures, we have previously mentioned that the work of Feracci *et al*. [22] deposited several PDB data, with the following accession numbers and descriptions:

-5EL3: Structure of the KH domain of T-STAR

-5ELR: Structure of the KH-QUA2 domain of T-STAR in complex with AAUAAU RNA

-5ELS: Structure of the KH domain of T-STAR in complex with AAAUAA RNA

-5ELT: Structure of the QUA1-KH domain of T-STAR in complex with UAAU RNA

-5EMO: Structure of the star domain of T-STAR in complex with AUUAAA RNA

In order to study the current PDB structures available, Win*Coot* [52] and PyMOL [53] programs were used. With Win*Coot*, the electron density maps (EDM) of the different PDB are displayed and the model deposited can be studied and analyzed using different tools in the program. The EDM have two densities: one represented as a blue mesh (the 2Fo-Fc map), which depicts the actual electron density of the protein, and the green and red density (the Fo-Fc map); green and red blobs show either peaks in the map where something is modeled but there is not enough data to support it (the red blobs) or where the data suggest there must be something modeled but there is nothing in the model at this time (green blobs). The EDM were obtained either using the Uppsala Density Server [54] or using the mmCIF data available at PDB website and converting the files into .mtz files using CCP4i. Regions corresponding to the RNA-binding sites were analyzed. The DSSP Secondary Structure available in the PDB website was also used to view the secondary structure prediction [55].

2.3.1. 3QHE

This structure corresponds to the interaction of the Tyrosine-rich domain of Sam68 with the adenomatous polyposis coli protein [56]. This domain is located in the positions 365-419, although with no secondary structure assigned in this experiment (Figure 10).

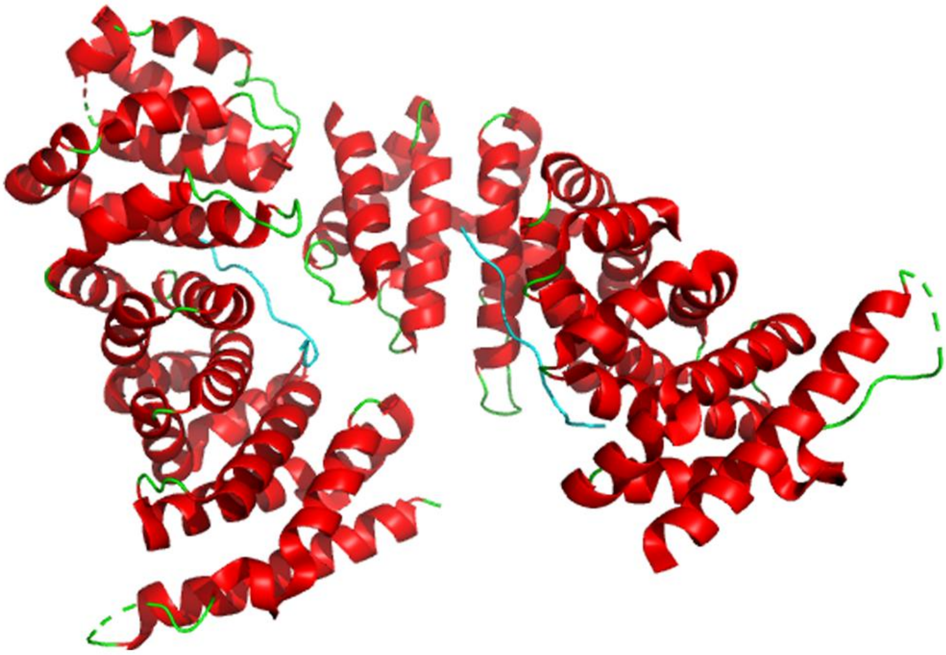
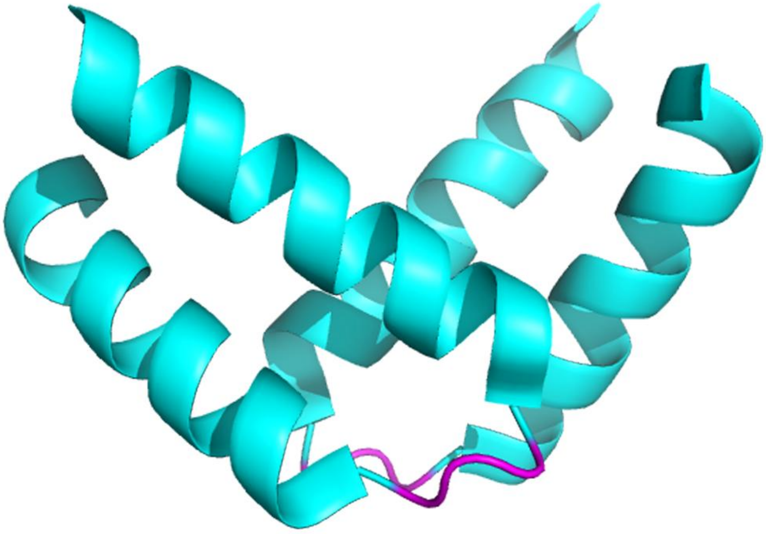


Figure 10. Schematic view of the PDB 3QHE. The tyrosine-rich domains of Sam68 are represented in cyan.

2.3.2. 2XA6

This structure represents the homodimerization of the QUA1 domain [57], which is next to the KH domain. It is located in positions 99-136, and it is mainly α-helix (73%). The schematic view of the secondary structure is shown in Figure 11.



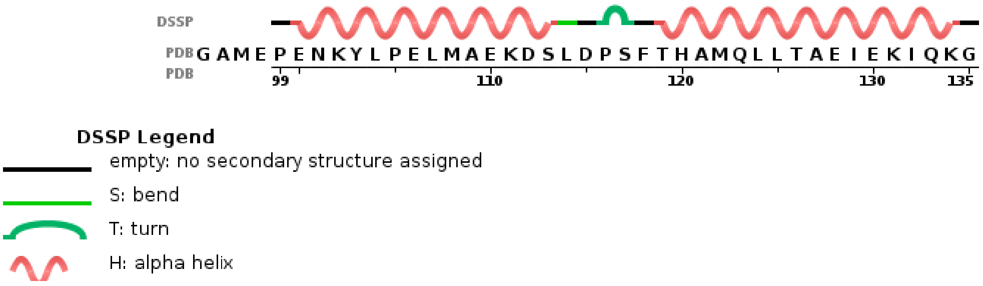
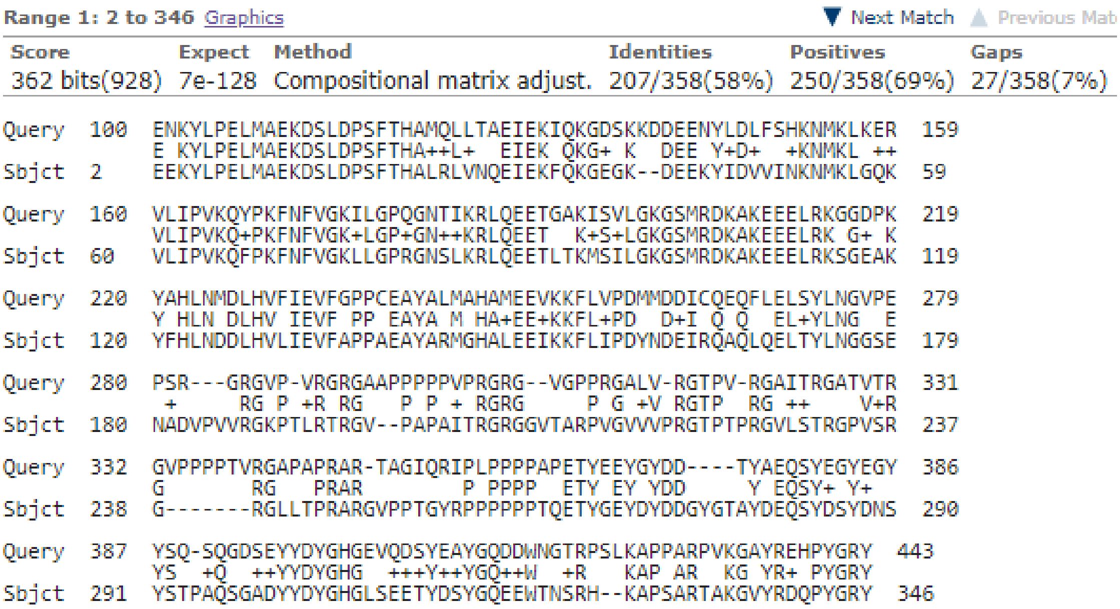

DSSP
PDB
GAMEPENKYLPELMAEKDSLDPSFTHAMQLLTAEIEKIQKG
PDB
99
110
120
130
135
DSSP Legend
empty: no secondary structure assigned
S: bend
T: turn
H: alpha helix

Figure 11. Schematic view of the PDB 2XA6.

2.3.3. Structures provided by Feracci *et al*.

The following structures were deposited by Feracci *et al*. [22]. They used T-STAR protein and its interaction with RNA instead of Sam68. But these proteins share similar KH domains, which is the site for RNA binding. A local alignment for these proteins was performed using the NCBI's BLASTp 2.8.0 (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins), with the first query Q07666 (Sam68) and the second query O75525 (T-STAR), and it is shown in Figure 12.



```
Range 1: 2 to 346

Score           Expect   Method                          Identities      Positives       Gaps
362 bits(928)   7e-128   Compositional matrix adjust.    207/358(58%)    250/358(69%)    27/358(7%)

Query  100  ENKYLPELMAEKDSLDPSFTHAMQLLTAEIEKIQKGDSKKDDEENYLDLFSHKNMKLKER  159
            E KYLPELMAEKDSLDPSFTHA++L+  EIEK QKG+ K  DEE Y+D+  +KNMKL ++
Sbjct  2    EEKYLPELMAEKDSLDPSFTHALRLVNQEIEKFQKGEGK--DEEKYIDVVINKNMKLGQK  59

Query  160  VLIPVKQYPKFNFVGKILGPQGNTIKRLQEETGAKISVLGKGSMRDKAKEEELRKGGDPK  219
            VLIPVKQ+PKFNFVGK+LGP+GN++KRLQEET  K+S+LGKGSMRDKAKEEELRK G+ K
Sbjct  60   VLIPVKQFPKFNFVGKLLGPRGNSLKRLQEETLTKMSILGKGSMRDKAKEEELRKSGEAK  119

Query  220  YAHLNMDLHVFIEVFGPPCEAYALMAHAMEEVKKFLVPDMMDDICQEQFLELSYLNGVPE  279
            Y HLN DLHV IEVF PP EAYA M HA+EE+KKFL+PD  D+I Q Q  EL+YLNG  E
Sbjct  120  YFHLNDDLHVLIEVFAPPAEAYARMGHALEEIKKFLIPDYNDEIRQAQLQELTYLNGGSE  179

Query  280  PSR---GRGVP-VRGRGAAPPPPPVPRGRG--VGPPRGALV-RGTPV-RGAITRGATVTR  331
             +     RG P +R RG   P P + RGRG     P G +V RGTP  RG ++    V+R
Sbjct  180  NADVPVVRGKPTLRTRGV--PAPAITRGRGGVTARPVGVVVPRGTPTPRGVLSTRGPVSR  237

Query  332  GVPPPPTVRGAPAPRAR-TAGIQRIPLPPPPAPETYEEYGYDD----TYAEQSYEGYEGY  386
            G       RG   PRAR        P PPPP  ETY EY YDD     Y EQSY+ Y+
Sbjct  238  G-------RGLLTPRARGVPPTGYRPPPPPPTQETYGEYDYDDGYGTAYDEQSYDSYDNS  290

Query  387  YSQ-SQGDSEYYDYGHGEVQDSYEAYGQDDWNGTRPSLKAPPARPVKGAYREHPYGRY  443
            YS  +Q  ++YYDYGHG  +++Y++YGQ++W  +R   KAP AR  KG YR+ PYGRY
Sbjct  291  YSTPAQSGADYYDYGHGLSEETYDSYGQEEWTNSRH--KAPSARTAKGVYRDQPYGRY  346
```

Figure 12. Local alignment of the proteins Sam68 (indicated as “Query”) and T-STAR (indicated as “Sbjct”). Both proteins share a 70% similarity.

2.3.3.1. 5EL3

This structure corresponds to the KH domain of the T-STAR, located in the positions 53-160. Figure 13 shows the 3D structure of the domain and the sequence chain view, characterized by a 44% of α-helix and 18% of β-sheet.

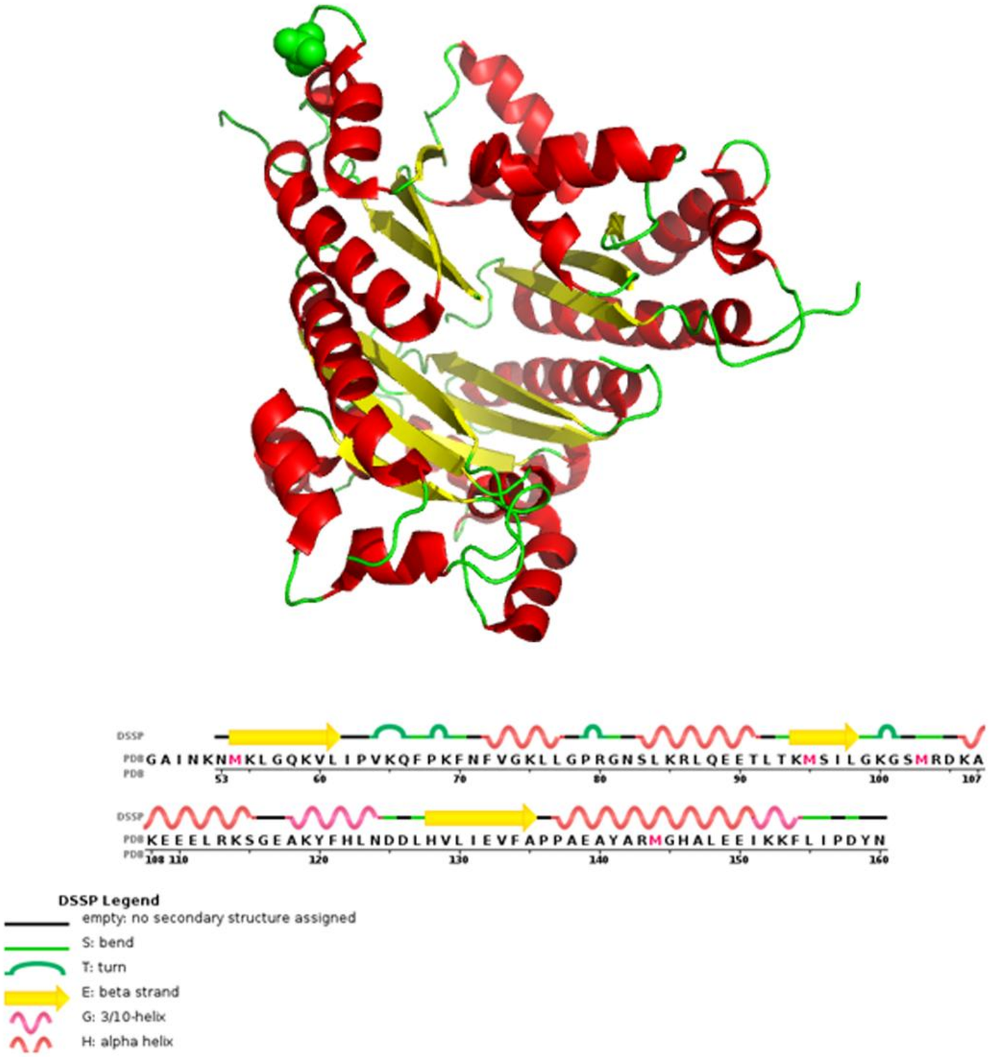

DSSP
PDB GAINKNMKLGQKVLIPVKQFPKFNFVGKLLGPRGNSLKRLQEETLTKMSILGKGSMRDKA
PDB 53 60 70 80 90 100 107
DSSP
PDB KEEELRKSGEAKYFHLNDDLHVLIEVFAPPAEAYARMGHALEEIKKFLIPDYN
PDB 108 110 120 130 140 150 160
DSSP Legend
empty: no secondary structure assigned
S: bend
T: turn
E: beta strand
G: 3/10-helix
H: alpha helix

Figure 13. View of the tridimensional structure of the KH domain of T-STAR protein and its secondary structure predicted by DSSP.

2.3.3.2. 5ELR

This PDB corresponds to the structure of the KH-QUA2 domain of T-STAR in complex with AAUAAU RNA, with a length of 136 residues and a 37% alpha-helix and 15% beta-sheet (Figure 14A). The PDB and UniProt sequences differ at two residue positions (48 and 49). The protein-RNA contact sites were consulted using PDBsum (www.ebi.ac.uk/pdbsum) and are shown in Figure 14B; a view of the PDB structure with the schematic representation of its secondary structure is shown in Figure 14C.

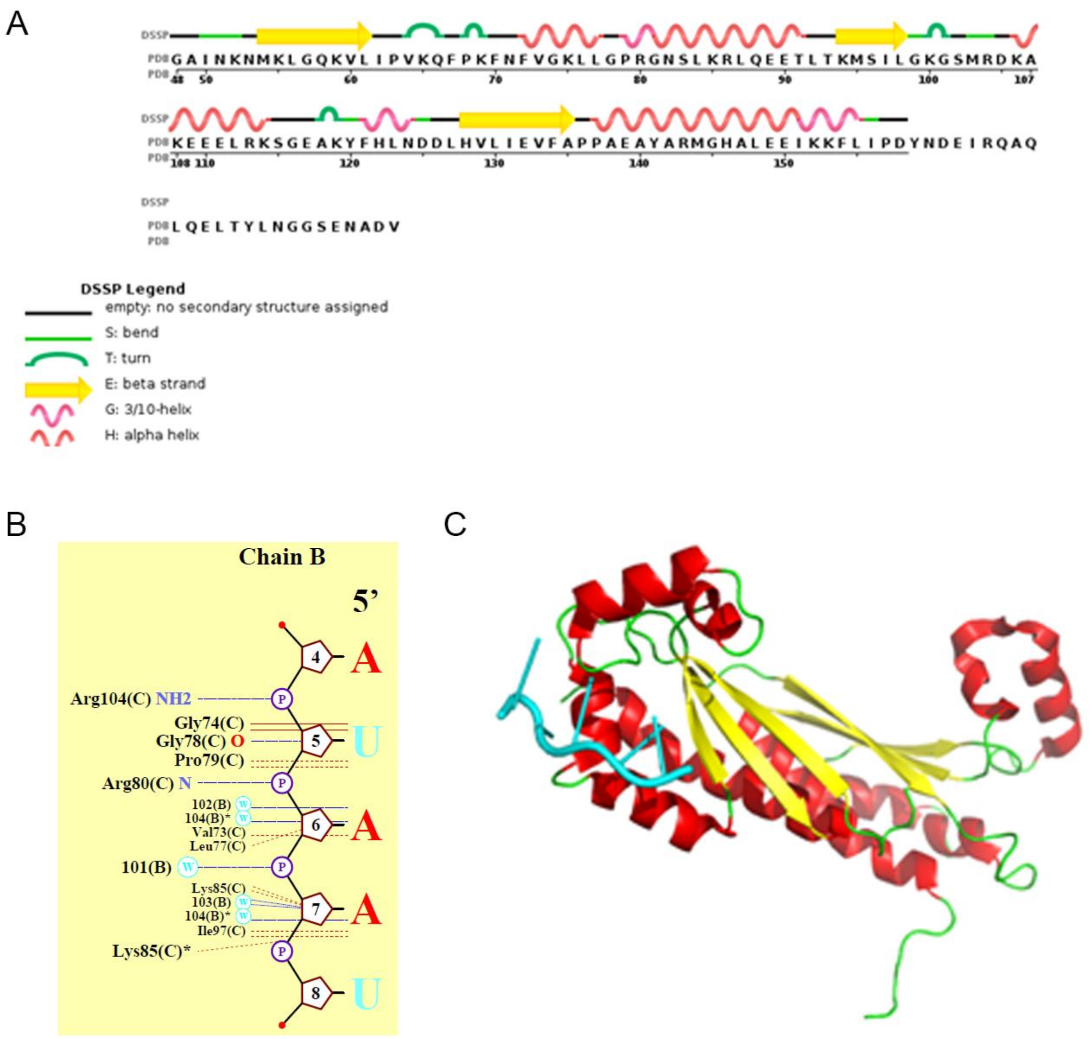


Figure 14. (A) Secondary structure of 5ELR predicted by DSSP. (B) Protein-RNA contact sites. The pentagon shows backbone sugar and base-number, the P shows the phosphate group; the star indicates residue/water on plot more than once, the blue dashes show hydrogen bonds, and the red dashes show non-bonded contact to DNA/RNA (< 3.35Å).(C) Schematic view of the KH-QUA2 domain, with the RNA motif showed in cyan.

Next, the model was analyzed using Win*Coot* with special attention to the regions where the protein-RNA interactions are located. Overall, the scores obtained in the density fit analysis are good in the residues in contact with the RNA (Figure 15); the lowest score was in the Arg80.

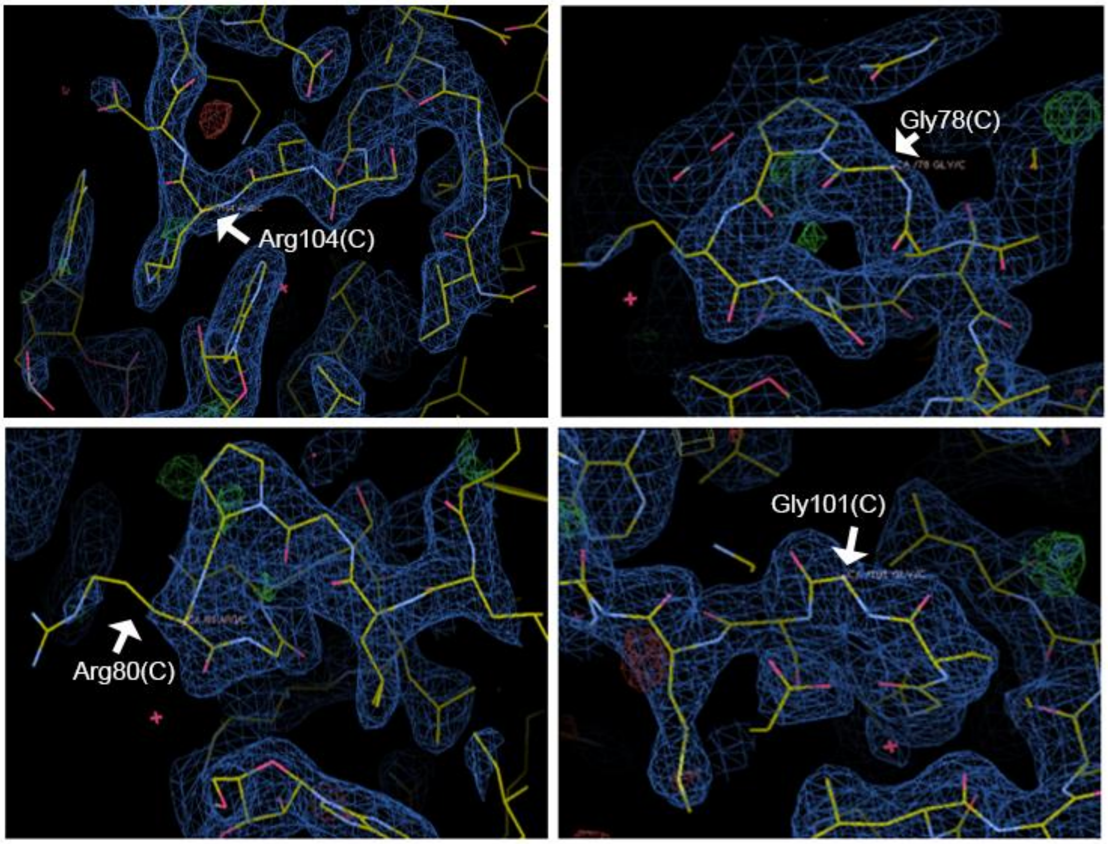


Figure 15. Some of the residues in contact with RNA and the EDM of the PDB 5ELR.

2.3.3.3. 5ELS

This structure consists of the KH-domain interacting with the RNA motif AAAUAA. Figure 16A shows the 3D structure and the secondary structure (46% helix, 18% beta-sheet), whereas the residues in contact with the RNA motif are shown in Figure 16B. Residues at positions 48 and 49 in the PDB differ from the sequence in UniProt. A view of the PDB structure with the schematic representation of its secondary structure is shown in Figure 16C.

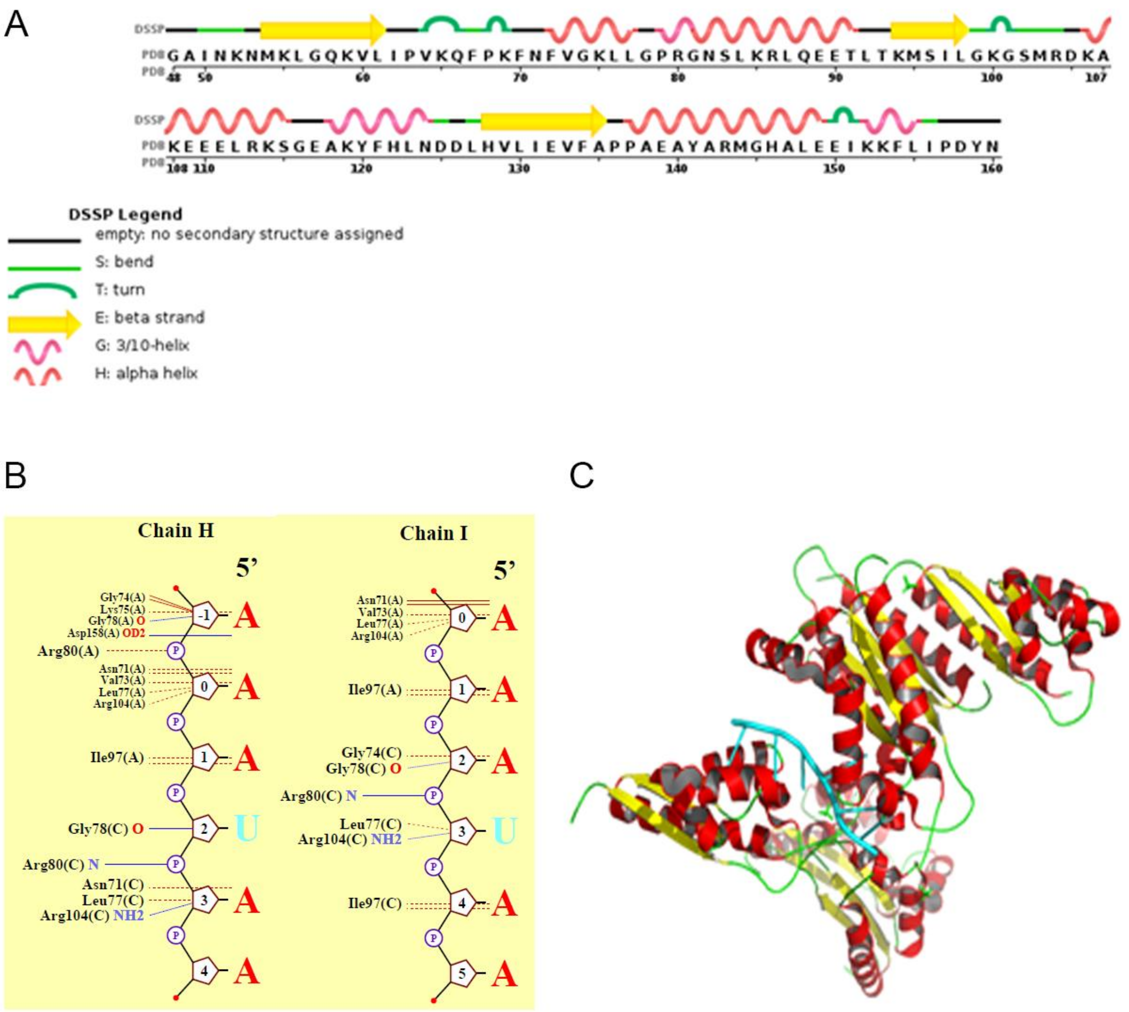


Figure 16. (A) Secondary structure of PDB 5ELS predicted by DSSP. (B) Protein-RNA contact sites. The pentagon shows backbone sugar and base-number, the P shows the phosphate group; the star indicates residue/water on plot more than once, the blue dashes show hydrogen bonds, and the red dashes show non-bonded contact to DNA/RNA (< 3.35Å). (C) Schematic view of the KH domain, with the RNA motif showed in cyan.

This structure contains 6 chains of the KH domain (A-F). The density fit graphs shown below (Figure 17) indicate that the scores worsen in chains D, E, and F, in comparison with chains A-C. However, the residues in contact with ARN have good scores, except for the Arg80 in chain C.

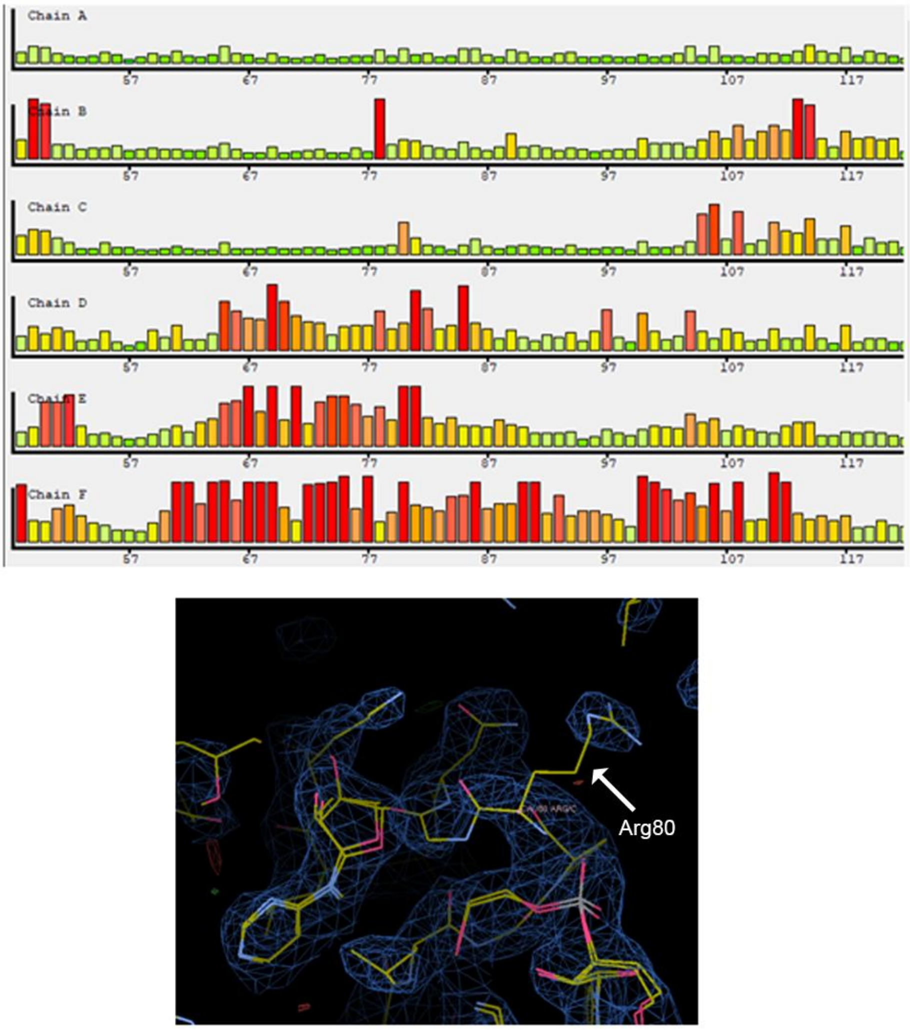


Figure 17. Density fit graphs for the EDM of 5ELS and a detail of the residue Arg80 in chain C.

2.3.3.4. 5ELT

This PDB represents the interaction of QUA1-KH domain with the RNA motif UAAU and the secondary structure predicted (46% helical, 14% beta-sheets; Figure 18A). The PDB sequence lacks in this case of the residues located in positions 39-44. PDBsum was used to know the residues where the protein-RNA interaction takes place, which is shown in Figure 18B. The PDB structure is shown in Figure 18C.

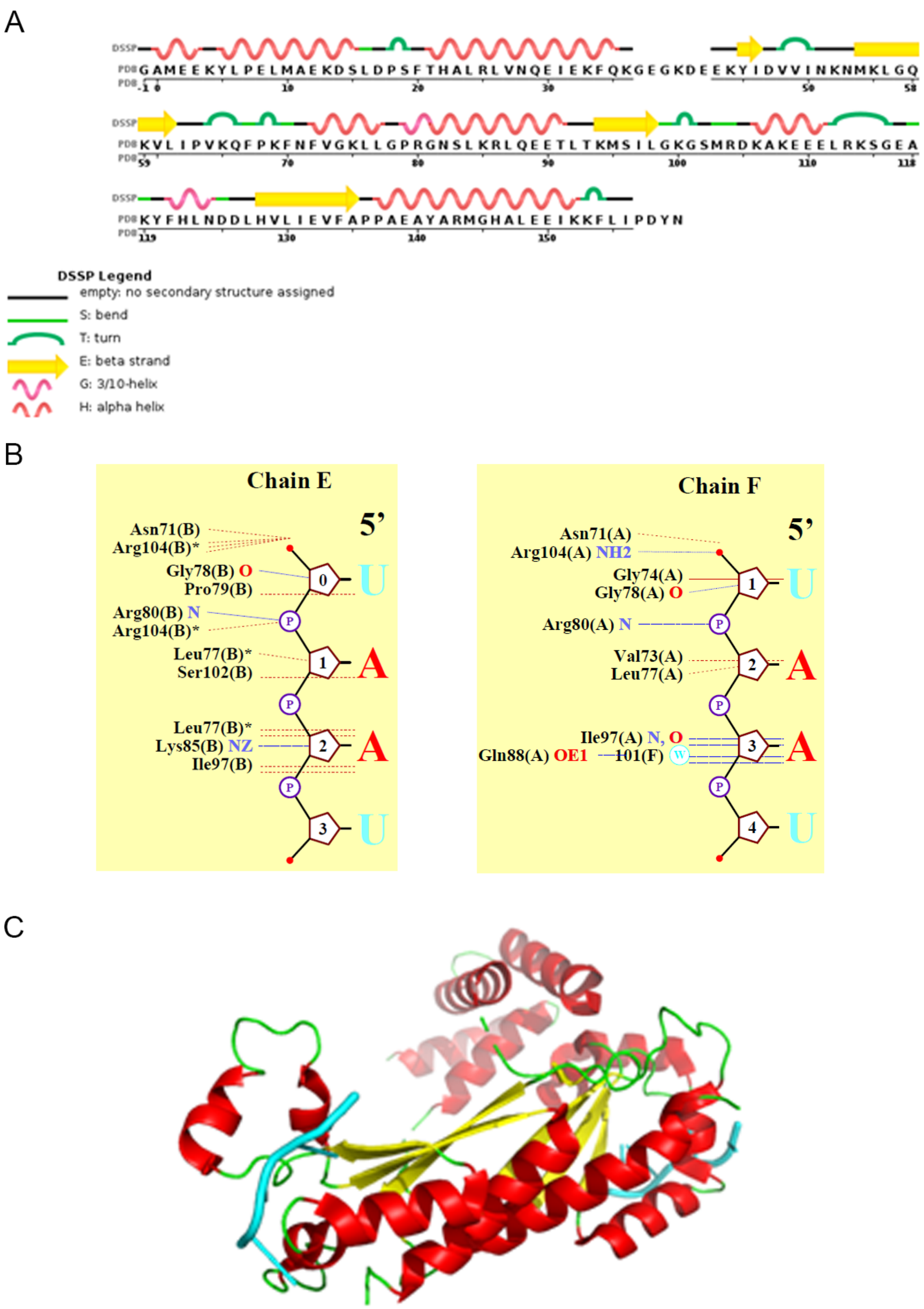


Figure 18. (A) Secondary structure of PDB 5ELT predicted by DSSP. (B) Protein-RNA contact sites. The pentagon shows backbone sugar and base-number, the P shows the phosphate group; the star indicates residue/water on plot more than once, the blue dashes show hydrogen bonds, and the red dashes show non-bonded contact to DNA/RNA (< 3.35Å). (C) Schematic view of the KH domain, with the RNA motif showed in cyan.

2.3.3.5. 5EMO

This structure represents the star domain (QUA1-KH-QUA2) of the T-STAR protein in complex with an AUUAAA RNA motif, and its secondary structure predicted by DSSP is shown in Figure 19A (37% alpha-helix and 11% beta-sheet). The PDB sequence lacks nine residues at positions 36-44 in comparison with the sequence in UniProt. The RNA-protein contact sites were determined using PDBsum (Figure 19B), and an overview of the tridimensional structure is shown in Figure 19C.

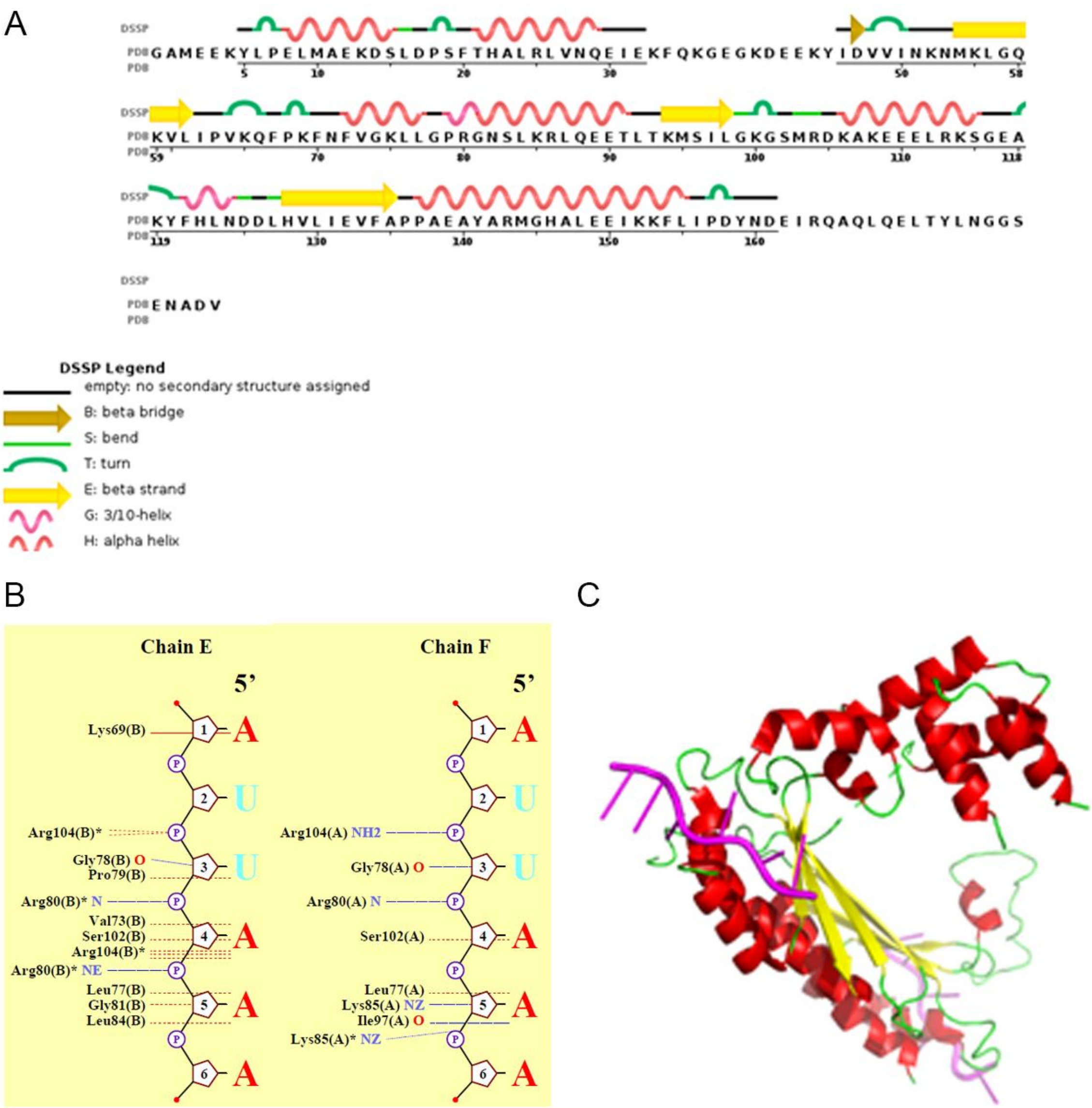


Figure 19. (A) Secondary structure of PDB 5EMO predicted by DSSP. (B) Protein-RNA contact sites. The pentagon shows backbone sugar and base-number, the P shows the phosphate group; the star indicates residue/water on plot more than once, the blue dashes show hydrogen bonds, and the red dashes show non-bonded contact to DNA/RNA (< 3.35Å). (C) Schematic view of the star domain, with the RNA motif showed in magenta.

Then the RNA-protein contact sites where checked using the EDM in Win*Coot*, and some of them are represented in Figure 20. It is worth mentioning that the lowest scores obtained in the density fit analysis were in the Arg80, as in the other PDB structures (5ELR and 5ELT). Lys85 in chain A had also a low score, whereas the residue Gly78 in chain B had better score.

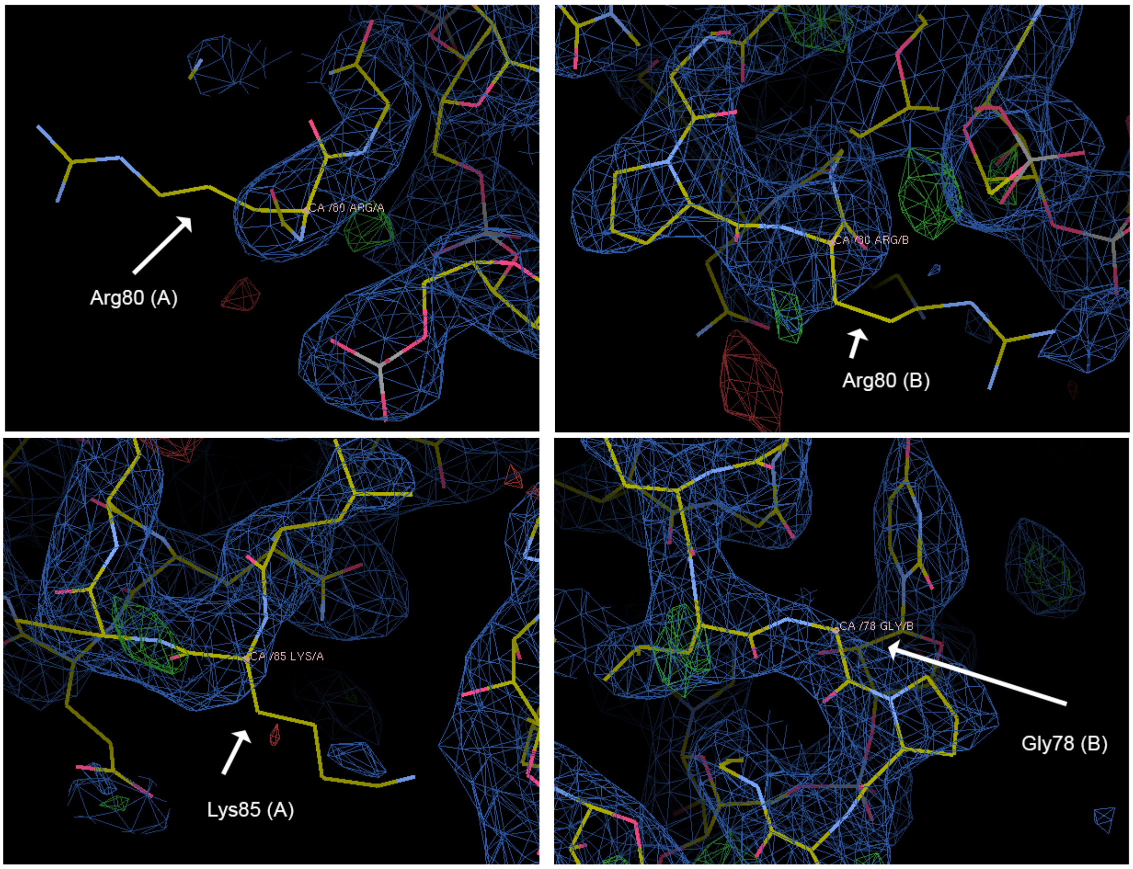


Figure 20. Some of the residues in contact with RNA and the EDM of the PDB 5EMO. The identity of the chain (A or B) is specified in brackets.

2.3.4. Summary of the study of PDB structures

All these PDB structures presented above provide a large amount of information about the structure of Sam68, since most of them are focused on the KH domain, which is the RNA-binding domain and therefore, a region of interest. The PDB sequences of 3QHE and 2XA6 come from Sam68 sequence (Figure 21), whereas the rest of the PDB structures use T-STAR sequence, which shares around 70% identity with Sam68. These data pave the way for the prediction of the structure of Sam68. Figure 22 shows the alignment between these sequences and Sam68 sequence.

```
>3qhe_B mol: protein length:55  KH domain-containing, RNA-binding,
signal transduction-associated protein 1
          Length = 55
 Score =  127 bits (318), Expect = 1e-28
 Identities = 55/55 (100%), Positives = 55/55 (100%)
Query: 365 TYEEYGYDDTYAEQSYEGYEGYYSQSQGDSEYYDYGHGEVQDSYEAYGQDDWNGT 419
           TYEEYGYDDTYAEQSYEGYEGYYSQSQGDSEYYDYGHGEVQDSYEAYGQDDWNGT
Sbjct: 1   TYEEYGYDDTYAEQSYEGYEGYYSQSQGDSEYYDYGHGEVQDSYEAYGQDDWNGT 55

>2xa6_B   mol:   protein   length:41      KH   DOMAIN-CONTAINING,RNA-
BINDING,SIGNAL TRANSDUCTION-ASSOCIATED PROTEIN 1
          Length = 41
 Score = 81.3 bits (199), Expect = 7e-15
 Identities = 39/39 (100%), Positives = 39/39 (100%)
Query: 97  MEPENKYLPELMAEKDSLDPSFTHAMQLLTAEIEKIQKG 135
           MEPENKYLPELMAEKDSLDPSFTHAMQLLTAEIEKIQKG
Sbjct: 3   MEPENKYLPELMAEKDSLDPSFTHAMQLLTAEIEKIQKG 41
```

Figure 21. Sequence alignment between Sam68 (Query) and the PDB structures (Sbjct). Only one chain is shown.

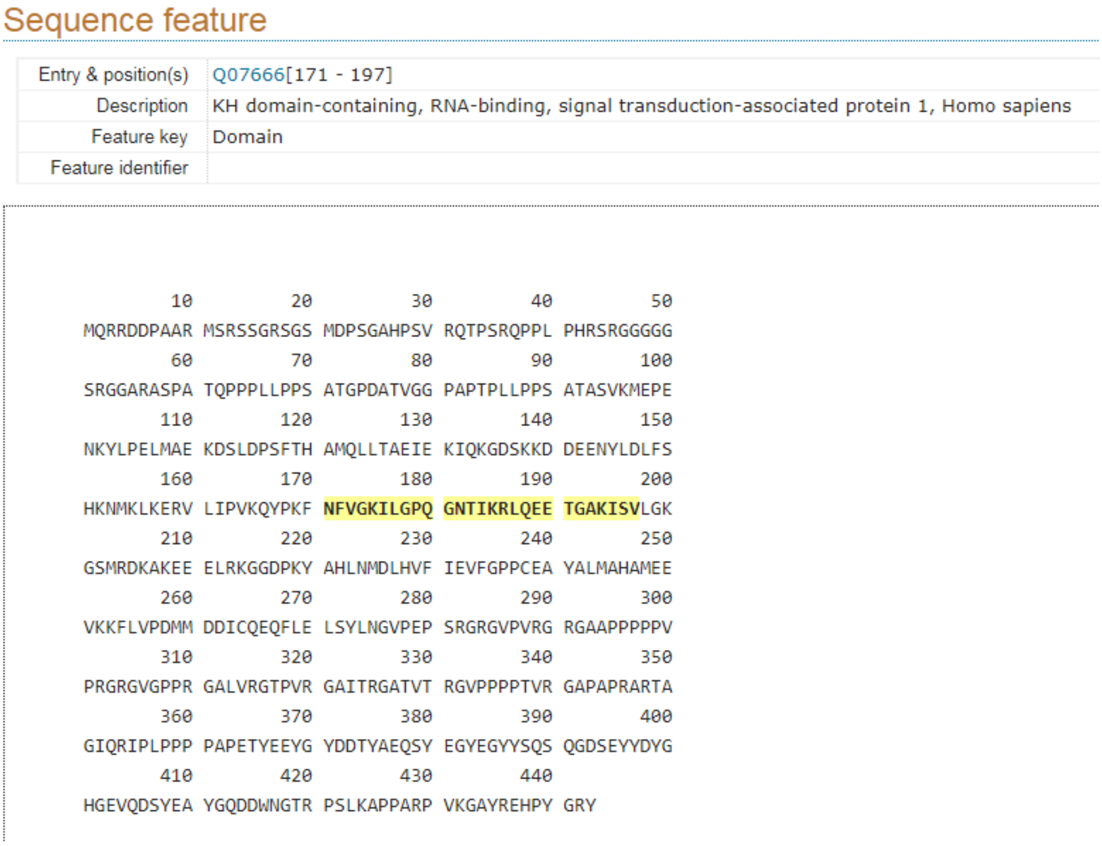
```
>5emo_B mol: protein length:185  KH domain-containing, RNA-binding,
signal transduction-associated protein 3
          Length = 185
 Score =  261 bits (666), Expect = 5e-69
 Identities = 128/180 (71%), Positives = 150/180 (83%), Gaps = 2/180
(1%)
Query:
100 ENKYLPELMAEKDSLDPSFTHAMQLLTAEIEKIQKGDSKKDDEENYLDLFSHKNMKLKER 159
    E KYLPELMAEKDSLDPSFTHA++L+  EIEK QKG+ K  DEE Y+D+  +KNMKL ++
Sbjct:
4   EEKYLPELMAEKDSLDPSFTHALRLVNQEIEKFQKGEGK--DEEKYIDVVINKNMKLGQK 61
Query:
160 VLIPVKQYPKFNFVGKILGPQGNTIKRLQEETGAKISVLGKGSMRDKAKEEELRKGGDPK 219
    VLIPVKQ+PKFNFVGK+LGP+GN++KRLQEET  K+S+LGKGSMRDKAKEEELRK G+ K
Sbjct:
62  VLIPVKQFPKFNFVGKLLGPRGNSLKRLQEETLTKMSILGKGSMRDKAKEEELRKSGEAK 121
Query:
220 YAHLNMDLHVFIEVFGPPCEAYALMAHAMEEVKKFLVPDMMDDICQEQFLELSYLNGVPE 279
    Y HLN DLHV IEVF PP EAYA M HA+EE+KKFL+PD  D+I Q Q  EL+YLNG  E
Sbjct:
122 YFHLNDDLHVLIEVFAPPAEAYARMGHALEEIKKFLIPDYNDEIRQAQLQELTYLNGGSE 181

>5elt_A mol: protein length:162  KH domain-containing, RNA-binding,
signal transduction-associated protein 3
          Length = 162
 Score =  243 bits (619), Expect = 1e-63
 Identities = 117/159 (73%), Positives = 137/159 (86%), Gaps = 2/159
(1%)
Query:
100 ENKYLPELMAEKDSLDPSFTHAMQLLTAEIEKIQKGDSKKDDEENYLDLFSHKNMKLKER 159
    E KYLPELMAEKDSLDPSFTHA++L+  EIEK QKG+ K  DEE Y+D+  +KNMKL ++
Sbjct:
4   EEKYLPELMAEKDSLDPSFTHALRLVNQEIEKFQKGEGK--DEEKYIDVVINKNMKLGQK 61
Query:
160 VLIPVKQYPKFNFVGKILGPQGNTIKRLQEETGAKISVLGKGSMRDKAKEEELRKGGDPK 219
    VLIPVKQ+PKFNFVGK+LGP+GN++KRLQEET  K+S+LGKGSMRDKAKEEELRK G+ K
Sbjct:
62  VLIPVKQFPKFNFVGKLLGPRGNSLKRLQEETLTKMSILGKGSMRDKAKEEELRKSGEAK 121
Query: 220 YAHLNMDLHVFIEVFGPPCEAYALMAHAMEEVKKFLVPD 258
           Y HLN DLHV IEVF PP EAYA M HA+EE+KKFL+PD
```

```
Sbjct: 122 YFHLNDDLHVLIEVFAPPAEAYARMGHALEEIKKFLIPD 160

>5elr_D mol: protein length:136  KH domain-containing, RNA-binding,
signal transduction-associated protein 3
          Length = 136
 Score =  193 bits (491), Expect = 9e-49
 Identities = 93/129 (72%), Positives = 109/129 (84%)

Query:
151 HKNMKLKERVLIPVKQYPKFNFVGKILGPQGNTIKRLQEETGAKISVLGKGSMRDKAKEE 210
    +KNMKL ++VLIPVKQ+PKFNFVGK+LGP+GN++KRLQEET  K+S+LGKGSMRDKAKEE
Sbjct:
4   NKNMKLGQKVLIPVKQFPKFNFVGKLLGPRGNSLKRLQEETLTKMSILGKGSMRDKAKEE 63
Query:
211 ELRKGGDPKYAHLNMDLHVFIEVFGPPCEAYALMAHAMEEVKKFLVPDMMDDICQEQFLE 270
    ELRK G+ KY HLN DLHV IEVF PP EAYA M HA+EE+KKFL+PD  D+I Q Q  E
Sbjct:
64  ELRKSGEAKYFHLNDDLHVLIEVFAPPAEAYARMGHALEEIKKFLIPDYNDEIRQAQLQE 123
Query: 271 LSYLNGVPE 279
           L+YLNG  E
Sbjct: 124 LTYLNGGSE 132

>5els_F  mol:protein  length:113   KH  domain-containing,  RNA-binding,
signal transduction-associated protein 3
          Length = 113
 Score =  175 bits (444), Expect = 3e-43
 Identities = 82/108 (75%), Positives = 96/108 (88%)
Query:
151 HKNMKLKERVLIPVKQYPKFNFVGKILGPQGNTIKRLQEETGAKISVLGKGSMRDKAKEE 210
    +KNMKL ++VLIPVKQ+PKFNFVGK+LGP+GN++KRLQEET  K+S+LGKGSMRDKAKEE
Sbjct:
4   NKNMKLGQKVLIPVKQFPKFNFVGKLLGPRGNSLKRLQEETLTKMSILGKGSMRDKAKEE 63
Query: 211 ELRKGGDPKYAHLNMDLHVFIEVFGPPCEAYALMAHAMEEVKKFLVPD 258
           ELRK G+ KY HLN DLHV IEVF PP EAYA M HA+EE+KKFL+PD
Sbjct: 64  ELRKSGEAKYFHLNDDLHVLIEVFAPPAEAYARMGHALEEIKKFLIPD 111

>5el3_D mol: protein length:113  KH domain-containing, RNA-binding,
signal transduction-associated protein 3
          Length = 113
 Score =  175 bits (444), Expect = 3e-43
 Identities = 82/108 (75%), Positives = 96/108 (88%)
Query:
151 HKNMKLKERVLIPVKQYPKFNFVGKILGPQGNTIKRLQEETGAKISVLGKGSMRDKAKEE 210
    +KNMKL ++VLIPVKQ+PKFNFVGK+LGP+GN++KRLQEET  K+S+LGKGSMRDKAKEE
Sbjct:
4   NKNMKLGQKVLIPVKQFPKFNFVGKLLGPRGNSLKRLQEETLTKMSILGKGSMRDKAKEE 63
Query: 211 ELRKGGDPKYAHLNMDLHVFIEVFGPPCEAYALMAHAMEEVKKFLVPD 258
           ELRK G+ KY HLN DLHV IEVF PP EAYA M HA+EE+KKFL+PD
Sbjct: 64  ELRKSGEAKYFHLNDDLHVLIEVFAPPAEAYARMGHALEEIKKFLIPD 111
```

Figure 22. Sequence alignment between Sam68 (Query) and the PDB structures (Sbjct). Only one chain is shown.

2.4. Domains of Sam68

The Sam68 entry in UniProt also contains information about family and domains, where the KH can be found in positions 171-197 (Figure 23).



Sequence feature

Entry & position(s)	Q07666[171 - 197]
Description	KH domain-containing, RNA-binding, signal transduction-associated protein 1, Homo sapiens
Feature key	Domain
Feature identifier	

```
        10         20         30         40         50
MQRRDDPAAR MSRSSGRSGS MDPSGAHPSV RQTPSRQPPL PHRSRGGGGG
        60         70         80         90        100
SRGGARASPA TQPPPLLPPS ATGPDATVGG PAPTPLLPPS ATASVKMEPE
       110        120        130        140        150
NKYLPELMAE KDSLDPSFTH AMQLLTAEIE KIQKGDSKKD DEENYLDLFS
       160        170        180        190        200
HKNMKLKERV LIPVKQYPKF NFVGKILGPQ GNTIKRLQEE TGAKISVLGK
       210        220        230        240        250
GSMRDKAKEE ELRKGGDPKY AHLNMDLHVF IEVFGPPCEA YALMAHAMEE
       260        270        280        290        300
VKKFLVPDMM DDICQEQFLE LSYLNGVPEP SRGRGVPVRG RGAAPPPPPV
       310        320        330        340        350
PRGRGVGPPR GALVRGTPVR GAITRGATVT RGVPPPPTVR GAPAPRARTA
       360        370        380        390        400
GIQRIPLPPP PAPETYEEYG YDDTYAEQSY EGYEGYYSQS QGDSEYYDYG
       410        420        430        440
HGEVQDSYEA YGQDDWNGTR PSLKAPPARP VKGAYREHPY GRY
```

Figure 23. Information about the domains of Sam68 in UniProt. This is the graph view of the KH domain, defined at 171-197 positions.

The InterPro database has also been consulted to identify protein families, domains, and functional sites. It identifies the homologous superfamily K Homology domain, subtype I (97 – 284 positions). Three domains are reported as well:

-Qua1 domain: 102 – 155

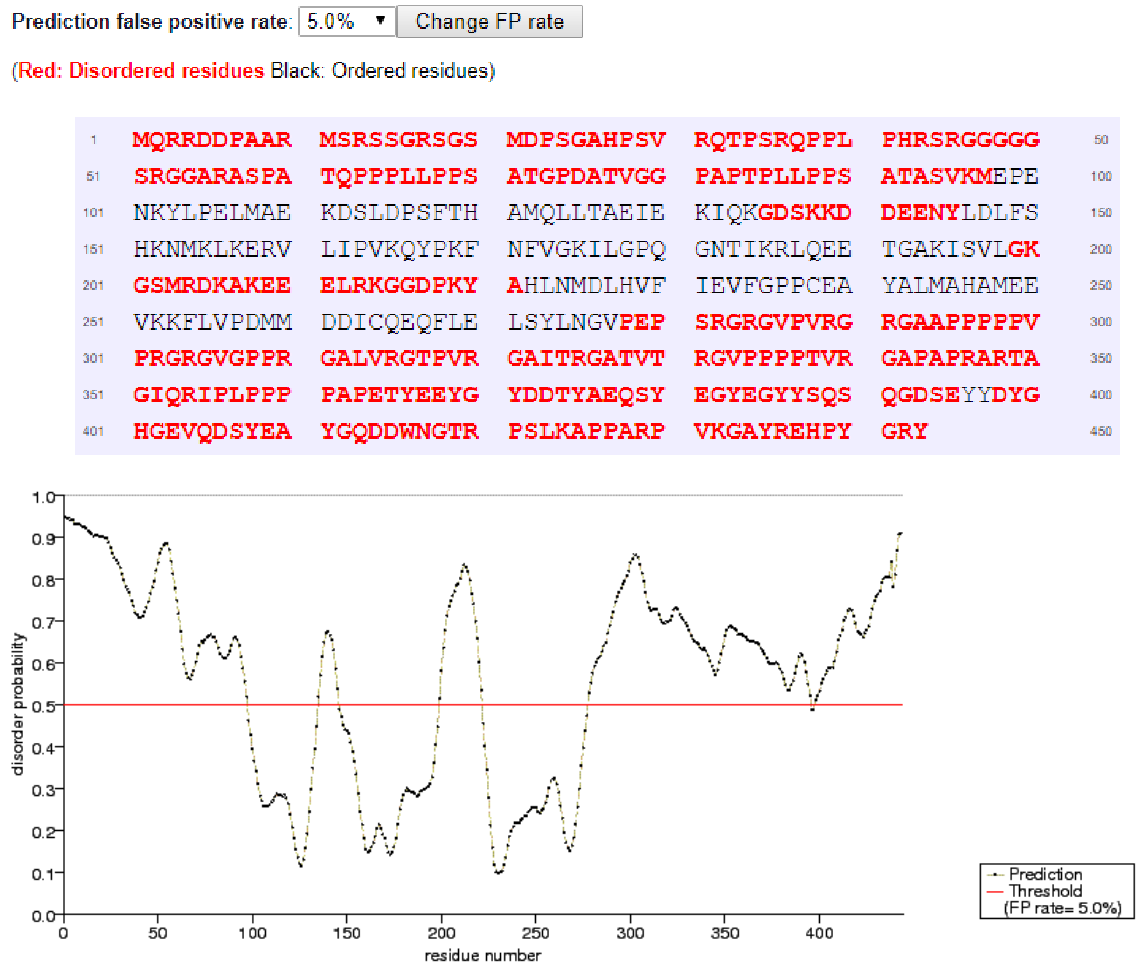
-K Homology domain, type I: 154 – 252

-Sam68, tyrosine rich domain: 366 – 415

3. Structure predictions of Sam68

3.1. Preliminary aspects

The well-known nuclear and cytoplasmic location of Sam68 has led us to rule out the study of transmembrane regions in its sequence. Disordered regions in the sequence have been analyzed using PrDOS server. The output is shown in Figure 24. Ordered residues are localized in 97 – 277 amino acids, approximately, which correspond to the GSG domain (which contains the KH domain) of Sam68. The rest of the residues of the sequence are considered disordered.



Prediction false positive rate: 5.0% ▼ Change FP rate

(Red: Disordered residues Black: Ordered residues)

```
  1  MQRRDDPAAR  MSRSSGRSGS  MDPSGAHPSV  RQTPSRQPPL  PHRSRGGGGG   50
 51  SRGGARASPA  TQPPPLLPPS  ATGPDATVGG  PAPTPLLPPS  ATASVKMEPE  100
101  NKYLPELMAE  KDSLDPSFTH  AMQLLTAEIE  KIQKGDSKKD  DEENYLDLFS  150
151  HKNMKLKERV  LIPVKQYPKF  NFVGKILGPQ  GNTIKRLQEE  TGAKISVLGK  200
201  GSMRDKAKEE  ELRKGGDPKY  AHLNMDLHVF  IEVFGPPCEA  YALMAHAMEE  250
251  VKKFLVPDMM  DDICQEQFLE  LSYLNGVPEP  SRGRGVPVRG  RGAAPPPPPV  300
301  PRGRGVGPPR  GALVRGTPVR  GAITRGATVT  RGVPPPPTVR  GAPAPRARTA  350
351  GIQRIPLPPP  PAPETYEEYG  YDDTYAEQSY  EGYEGYYSQS  QGDSEYYDYG  400
401  HGEVQDSYEA  YGQDDWNGTR  PSLKAPPARP  VKGAYREHPY  GRY         450
```

Figure 24. Output of the disordered residues in the sequence of Sam68.

In addition, the prediction of signal peptides was carried out using the SignalP 4.1. Server, although Sam68 sequence was unlikely to have a signal peptide. As shown in Figure 25, no signal peptide was identified.

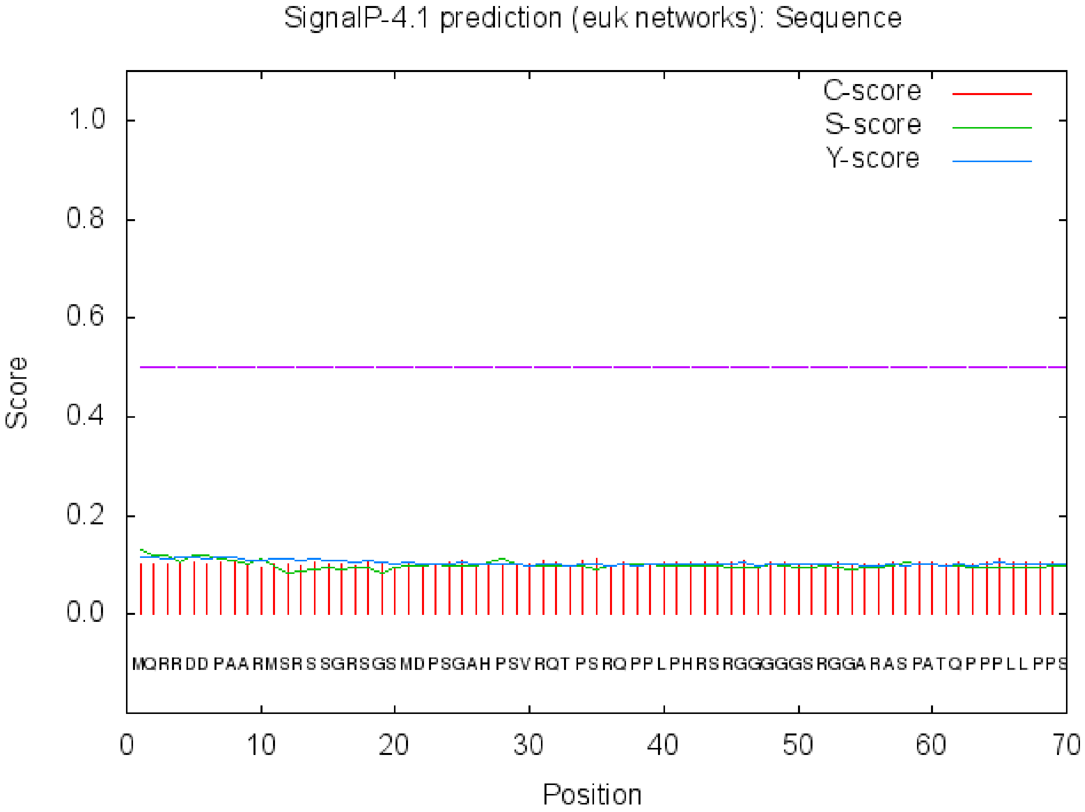
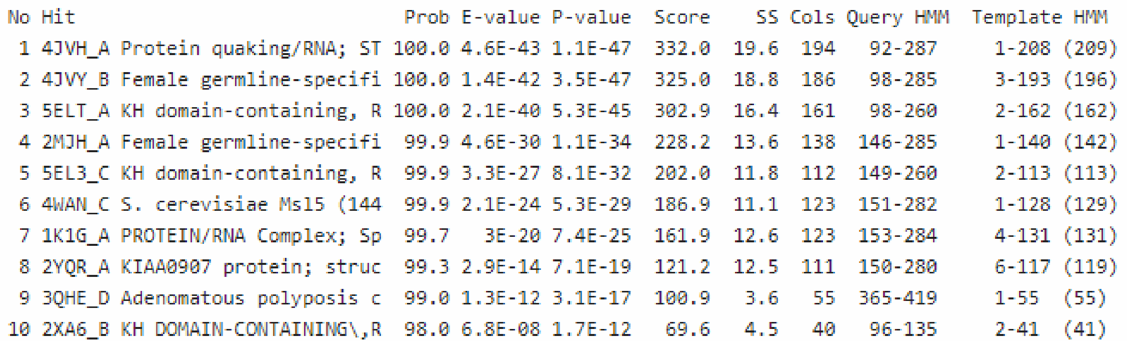


Figure 25. Output of the signal peptide analysis for the sequence of Sam68. The green lines represent the S-score, which is the signal peptide score.

3.2. Homology based predictions

3.2.1. HHPRED

HHPRED was used to search for templates for Sam68. It provides a list of homologous proteins and their alignments with the sequence of the target. Figure 26 shows the first ten hits of the output of this search, together with a schematic graph showing where the templates were found aligned with Sam68. The best templates found cover the GSG domain region (100 – 281 positions, approximately).



```
No Hit                             Prob E-value P-value  Score    SS Cols Query HMM  Template HMM
 1 4JVH_A Protein quaking/RNA; ST 100.0 4.6E-43 1.1E-47  332.0  19.6  194   92-287     1-208 (209)
 2 4JVY_B Female germline-specifi 100.0 1.4E-42 3.5E-47  325.0  18.8  186   98-285     3-193 (196)
 3 5ELT_A KH domain-containing, R 100.0 2.1E-40 5.3E-45  302.9  16.4  161   98-260     2-162 (162)
 4 2MJH_A Female germline-specifi  99.9 4.6E-30 1.1E-34  228.2  13.6  138  146-285     1-140 (142)
 5 5EL3_C KH domain-containing, R  99.9 3.3E-27 8.1E-32  202.0  11.8  112  149-260     2-113 (113)
 6 4WAN_C S. cerevisiae Msl5 (144  99.9 2.1E-24 5.3E-29  186.9  11.1  123  151-282     1-128 (129)
 7 1K1G_A PROTEIN/RNA Complex; Sp  99.7   3E-20 7.4E-25  161.9  12.6  123  153-284     4-131 (131)
 8 2YQR_A KIAA0907 protein; struc  99.3 2.9E-14 7.1E-19  121.2  12.5  111  150-280     6-117 (119)
 9 3QHE_D Adenomatous polyposis c  99.0 1.3E-12 3.1E-17  100.9   3.6   55  365-419     1-55  (55)
10 2XA6_B KH DOMAIN-CONTAINING\,R  98.0 6.8E-08 1.7E-12   69.6   4.5   40   96-135     2-41  (41)
```

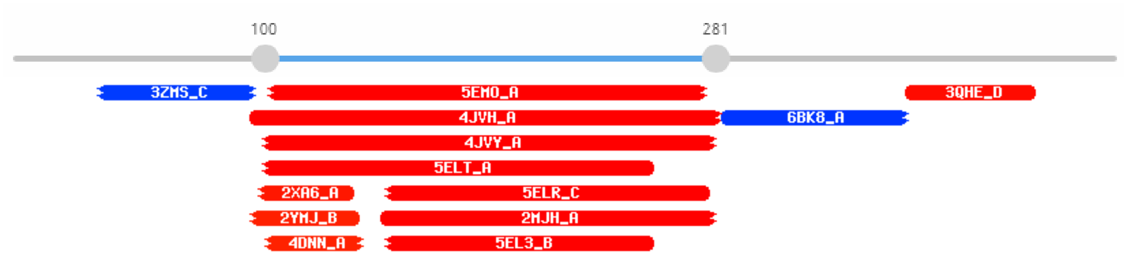


Figure 26. List of the closest homologous to the target Sam68. Prob indicates the true positive rate of the template; E-value provides the average number of false positives; the P-value is the result from dividing the E-value between the number of sequences; the Query HMM shows the region in the target sequence where the template is aligned, whereas the Template HMM indicates the region in the template where the target is aligned.

3.2.2. BhageerathH

This server provided five models, together with other assessments performed in order to make the prediction. They were found very similar in the five models proposed, therefore only the plots for Model 1 are shown.

- ProTSAV Quality Assessment of Query Structure (Figure 27): it is a server which evaluates the quality of a protein model using several assessment tools and shows the results in a color plot. Green color indicates the input structure to be in 0-2 Å rmsd (root-mean-square deviation), yellow color indicates 2-5 Å, orange color indicates 5-8 Å and red color indicates input structure to be beyond 8 Å. The overall quality score is shown as a dot.

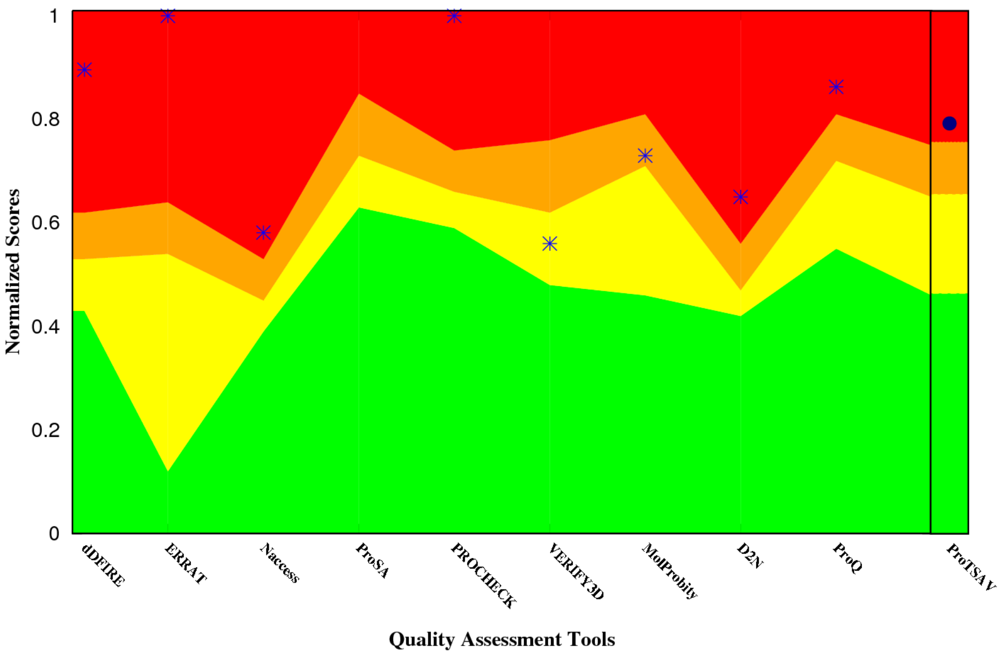


Figure 27. Dot plot with the scores for evaluating the quality of the model.

- NACCESS (Figure 28): This program calculates the accessibility of a molecule.

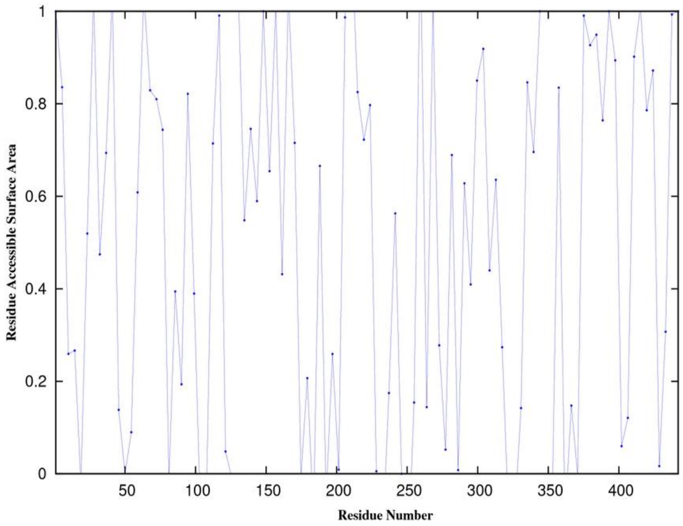


Figure 28. Accessibility of the residues represented by the accessible surface area.

- PROCHECK (Figure 29): Provides a Ramachandran plot; this diagram represents the torsion angles phi and psi, which represent the relative permitted rotation of the N-Cα and Cα-C bonds in the main chain in a protein, respectively. With this plot, energetically allowed regions for a residue, based on these torsion angles, can be visualized. The red regions represent the favored areas, mainly alpha-helix (lower left quadrant) and beta-sheet (upper left quadrant), whereas the yellow and light brown regions correspond to allowed and "generously allowed" regions, respectively. Upper right quadrant represents a less common conformation, the left-handed alpha-helix.

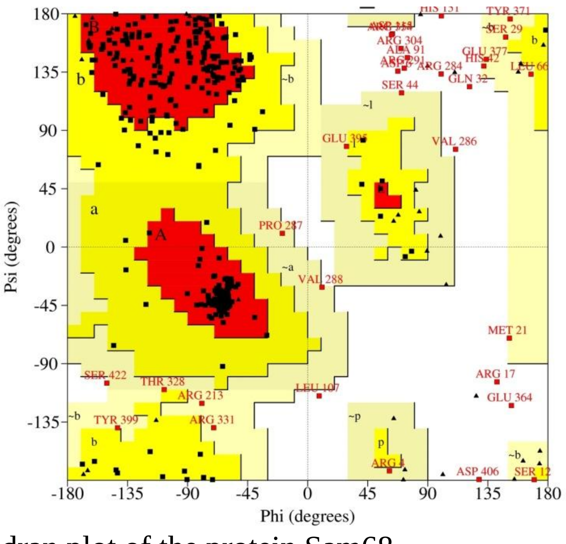


Figure 29. Ramachandran plot of the protein Sam68.

- VERIFY3D (Figure 30): This tool assesses the compatibility of the 3D model with the amino acid sequence (1D). It assigns a structural class based on its location and compares the results to good structures. The graph shown in Figure 30 represents the 3D-1D scores for each residue, which is a measure of the compatibility of a sequence with a protein structure based on three criteria: surface accessibility / area of the residue that is buried, the environment polarity (area covered by polar atoms), and the local secondary structure.

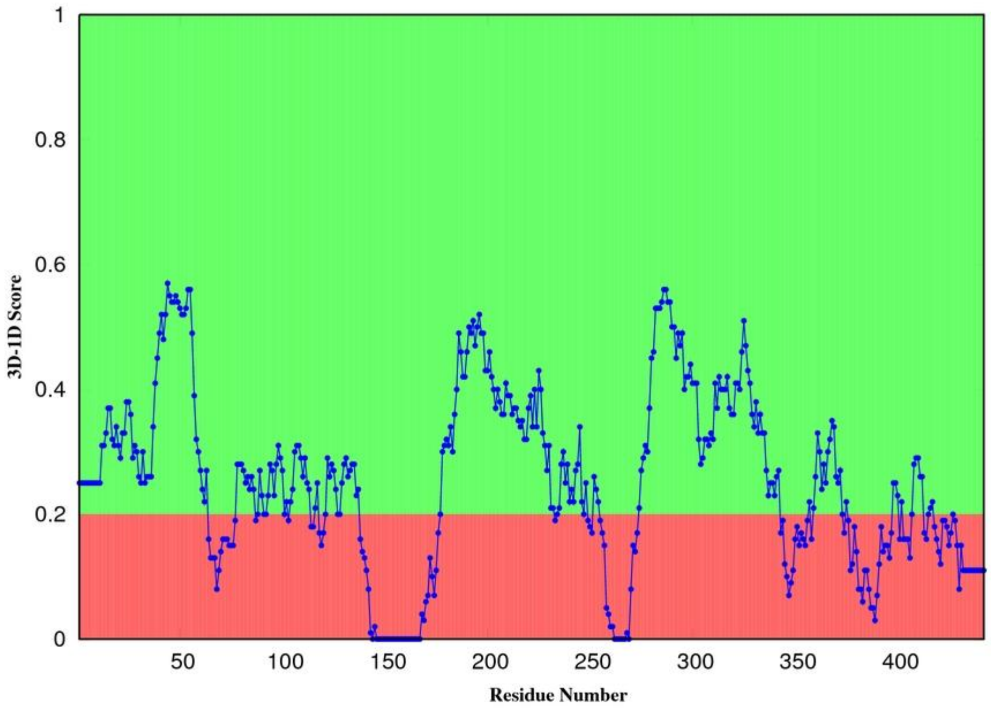


Figure 30. Dot plot representing the 3D-1D score. The highest scores are found in three different regions of the protein, including the KH domain.

Figure 31 shows the five models predicted by BhageerathH. The most significant helix regions are colored according to their order in the sequence. Three beta-sheets are predicted in the five models. The main difference between the models lies in the presence or absence of relative small alpha-helix structures (shown in red). The arrangement of the coil regions differs in the five models, mainly in the last part of the sequence. This can be due to the fact that this is a disordered region, which is not suitable for structure prediction.

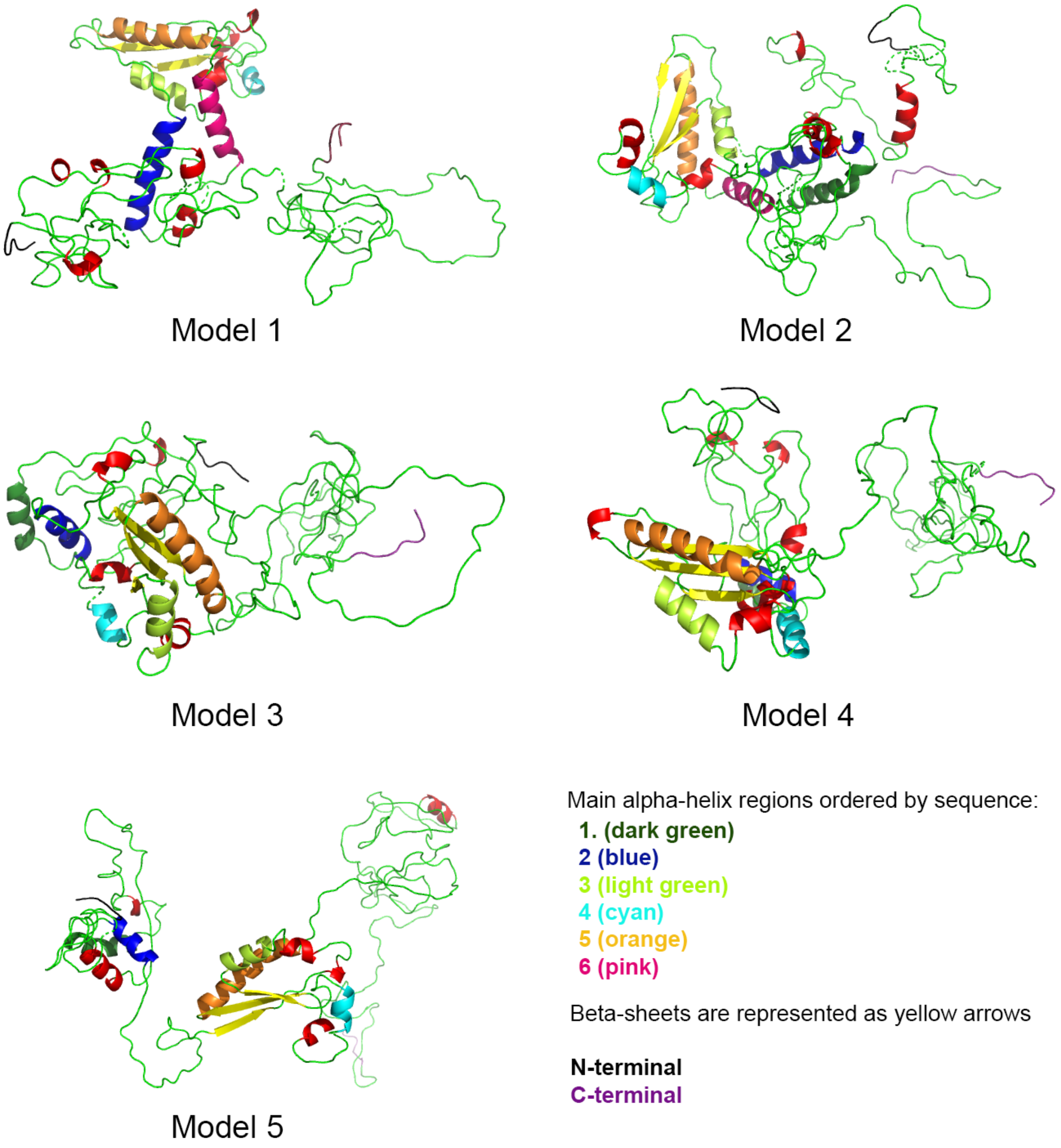


Figure 31. The five models predicted by BhageerathH server.

3.2.3. RAPTOR X

The model predicted by RAPTOR X (Figure 32) has used as templates the PDBs 5ELT-A and 5EMO-A, and comprises the 87 – 271 positions. It has an alignment score of 140 (out of the 443 residues). The unnormalized GDT (global distance test) score and the GDT score are used to estimate the model quality. In this case (a protein with >200 amino acids), an uGDT > 50 is a good indicator. The model predicted has an uGDT score of 128. The P-value of the 3D model is 3.25 x 10^{-6}, which is also a good indicator. These three measures (score, uGDT(GDT), and P-value) must be taken into account in order to judge the quality of the model. In this case, all the indicators point to a high quality model.

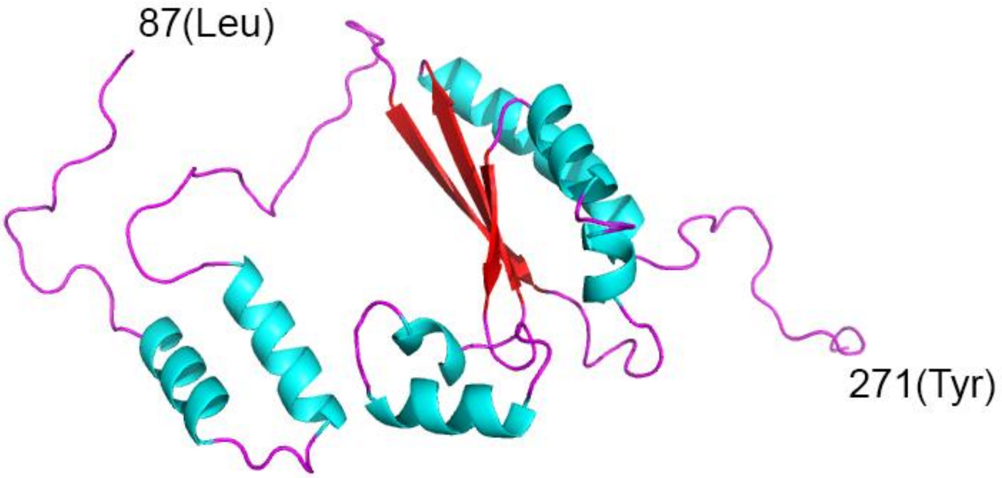


Figure 32. The 3D model of Sam68 predicted by RAPTOR X server (87 – 271 position). The helix regions are colored in cyan and the beta-sheets in red.

3.2.4. Phyre 2

The top five models are based on the templates with PDB accession numbers: 4JVY (chain A), 4JVH (chain A), 5ELT (chain B), 2BL5 (chain A1), and 4WAL (chain A). Figure 33 shows the model predicted by this server. The disordered regions shown as coils in purple are more compacted in comparison with the previous models, like surrounding the central domain.

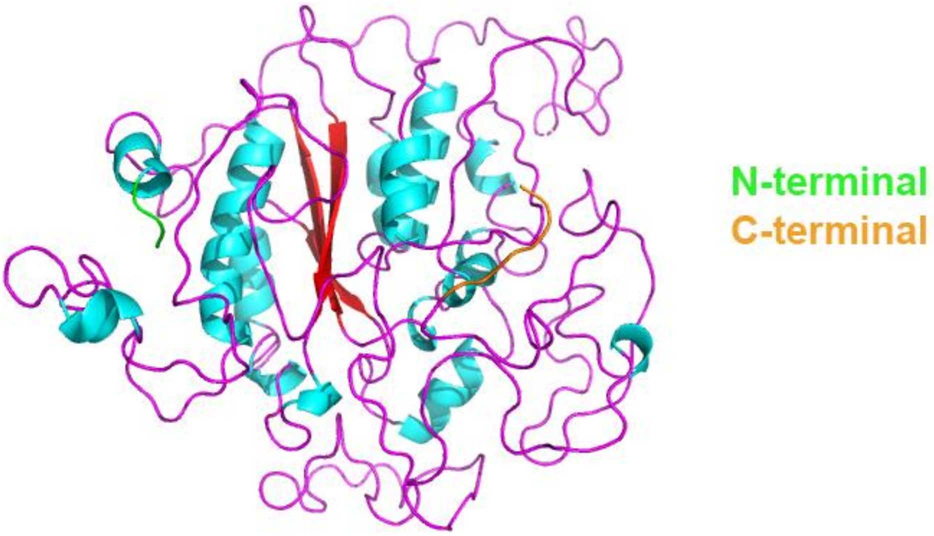


Figure 33. Predicted model of Sam68 by the Phyre2 server. The helix regions are colored in cyan and the beta-sheets in red. The N-terminal and C-terminal regions are highlighted in green and red, respectively.

3.2.5. IntFOLD3

The 3D structure predicted by IntFOLD3 is based on the PDB structure 4JVY-A. It has a global quality score of 0.3758, indicating that, although not incorrect, it is not a very confident model. This can be due to the fact that the individual residue errors (Figure 34) are high in those regions where no good templates can be found, thus lowering the global score. The confidence level assigned to this model is “medium”, which is the middle point between certain and poor, and indicates less than a 1/20 chance that the model is incorrect. Figure 35A shows the predicted model; part of it is not shown due to the lineal disposition of the initial and terminal regions, in order to better show the central domain.

Another 3D structure was predicted using IntFOLD server (Figure 35B), this time submitting the sequence of Sam68 which contains the GSG domain (81 – 280 positions). As expected, the global score improved in comparison with the previous model (0.665), and so did the p-value and the confidence level. The 3D structures of the sequence containing the GSG domain are very similar in both models.

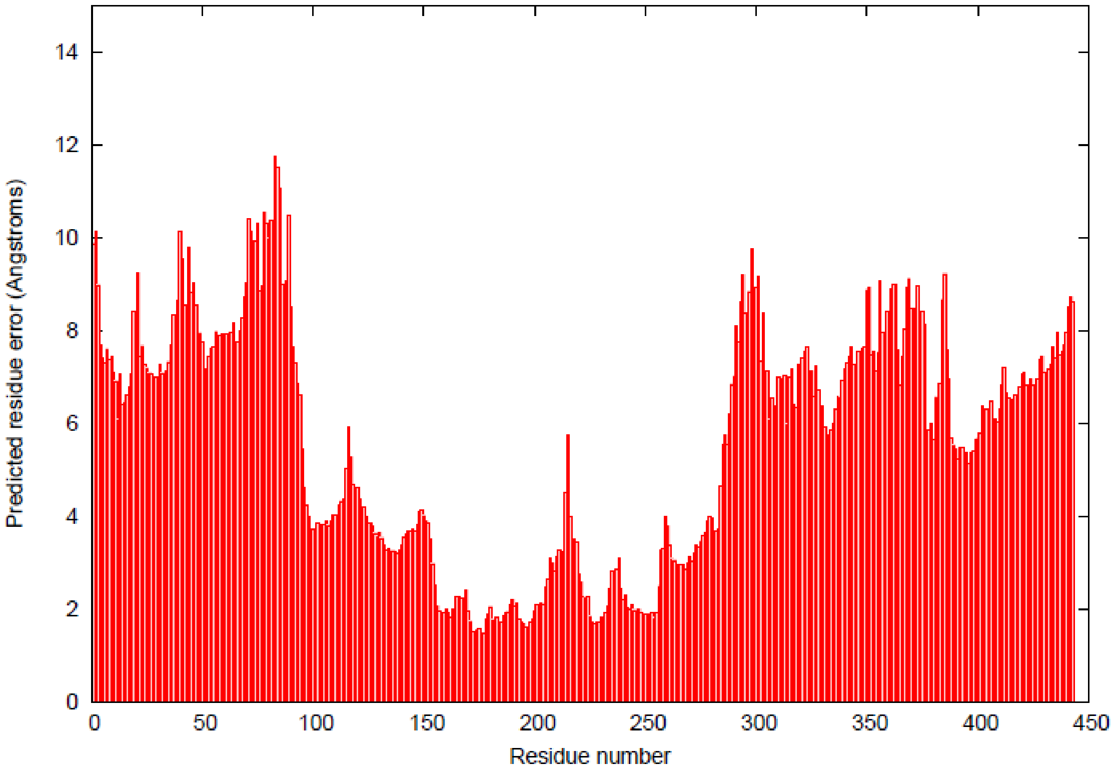


Figure 34. Plot of the predicted residue error provided by IntFOLD3 server.

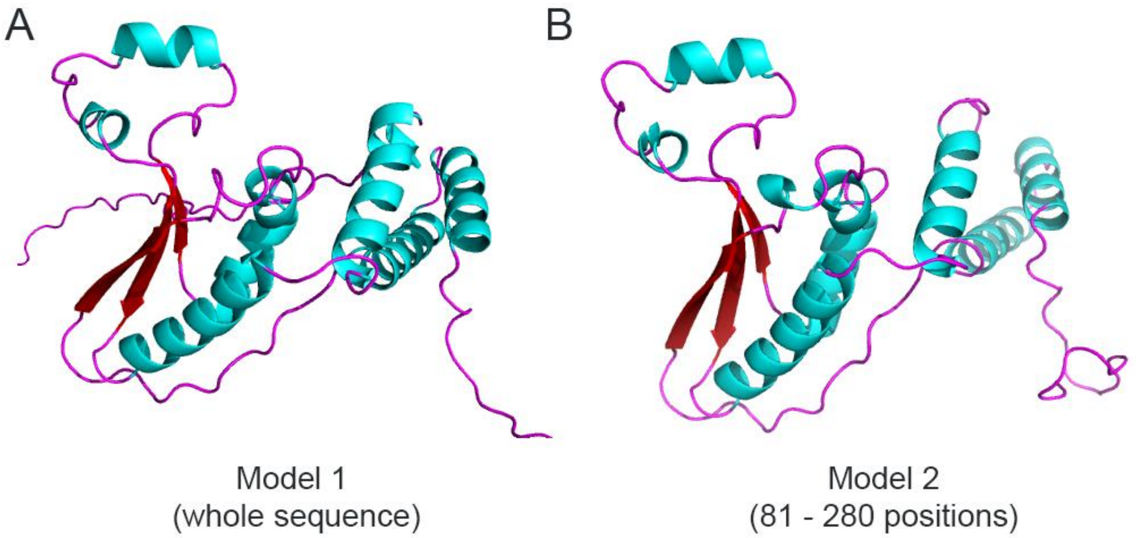


Figure 35. Predicted models by IntFOLD3. (A) Using the whole sequence of Sam68. (B) Using a region which contains the GSG (and the KH) domain of Sam68. The helix regions are colored in cyan and the beta-sheets in red.

3.2.6. SWISS-MODEL

The models predicted by SWISS-MODEL were based on the PDB structures: 5ELT-A, 4JVH-A, 5EMO-A. The global model quality estimate (GMQE) ranges between 0.25 and 0.28 (the score is expressed as a number between 0 and 1). As in the previous case, these low scores are due in part to the fact that the coverage of the templates is low in comparison with the total length of the sequence. Thus, the 3D structures provided cover the 98 – 280 region of the protein, as it can be shown in Figure 36.

3.2.7. Summary of the predicted models of Sam68

Some of the features that the models predicted by the different servers have in common are:

-The templates used by the servers were quite similar. Table 4 shows the PDB structures used as templates by the servers, except for BagheerathH, which did not provide this information.

-The global scores were low due to the fact that part of the sequence lacks of templates. Moreover, this area corresponds to a disordered region, where good predictions cannot be performed.

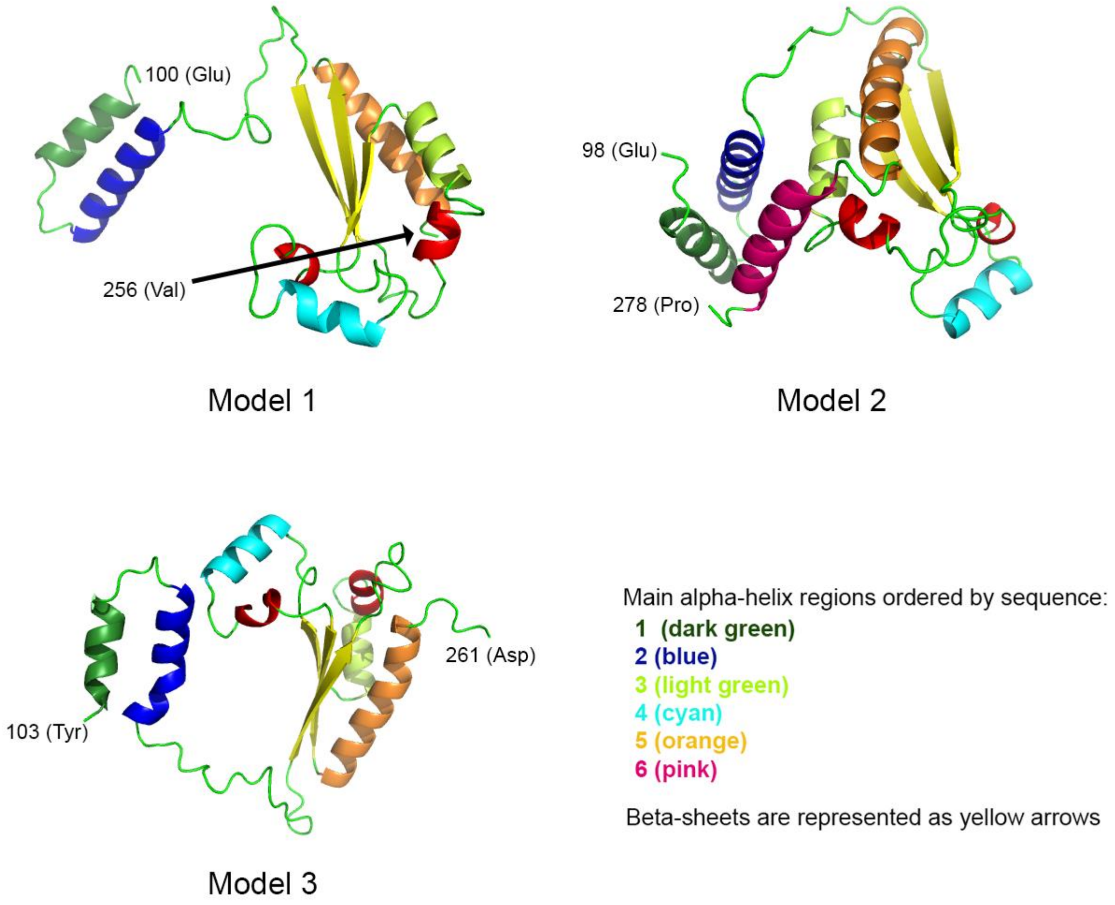


Figure 36. Models predicted by SWISS-MODEL.

	RAPTOR X	PHYRE 2	INTFOLD3	BHAGEERATH-H	SWISSMODEL
Templates	5ELT-A 5EMO-A	4JVY-A 4JVH-A 5ELT-B 2BL5-A1 4WAL-A	4JVY-A	Not available	5ELT-A 4JVH-A 5EMO-B

Table 4. Templates used for homology modeling by each server.

-The ~80 – 280 region is the best modeled area of the protein, where most of the secondary structures predicted are located. Table 5 shows a summary of the regions where alpha-helix and beta-sheet regions are located according to the different servers used.

	RAPTOR X	PHYRE 2	INTFOLD3	BHAGEERATH-H	SWISSMODEL
Helix	7 – 11	8 – 11		7 – 9	
		35 – 37			
		69 – 72			
		74 – 78			
	104 – 134	102 – 134	101 – 115	99 – 111	103 – 113
			119 – 137	117 – 133	119 – 134
			172 – 177		172 – 177
	183 – 191	181 – 191	182 – 191	183 – 191	182 – 191
	206 – 215		206 – 212	207 – 213	206 – 213
			220 – 224		221 – 224
	237 – 255	237 – 255	239 – 255	237 – 254	237 – 254
	261 – 275	262 – 275	262 – 276	260 – 274	
			279 – 281		
		346 – 350			
		404 – 407			
β-Sheet	155 – 163	155 – 161	154 – 161	156 – 162	154 – 161
	172 – 177	172 – 176		172 – 177	
	194 – 198	194 – 199	194 – 198	194 – 199	194 – 198
	228 – 234	227 – 232	228 – 235		228 – 235

Table 5. Summary of the positions within the Sam68 sequence, where the secondary structures alpha-helix and beta-sheet were predicted.

In general, the secondary structure prediction was quite similar in the five servers used. In the case of Phyre2 four alpha-helix are predicted in the beginning of the sequence (between 8 and 78 positions), whereas Raptor X and BhageerathH predict a small helix between 7 and 11 positions. The five servers predicted an alpha-helix in the region 99/104 – 133/137, although two of them divided this area into two helices connected by a loop. In the 181/183 – 191 and 237/239 – 254/255 positions an alpha-helix was predicted by the five servers as well. Four of them showed this secondary structure in 206/207 – 212/215 and 260/261 – 274/276 regions. Again, only Phyre2 predicted two small helices in the last part of the sequence of Sam68.

Regarding the beta-sheets, three or four were predicted and two of them were present in the five models: 154/156 – 161/163 and 194 – 198/199 regions. A beta-sheet in 227/228 – 232/235 was predicted in four out of the five servers consulted.

3.3. Protein threading

3.3.1. I-TASSER

This server predicted five models based on the compatibility analysis between the 3D structures and the linear protein sequence (Figure 37).

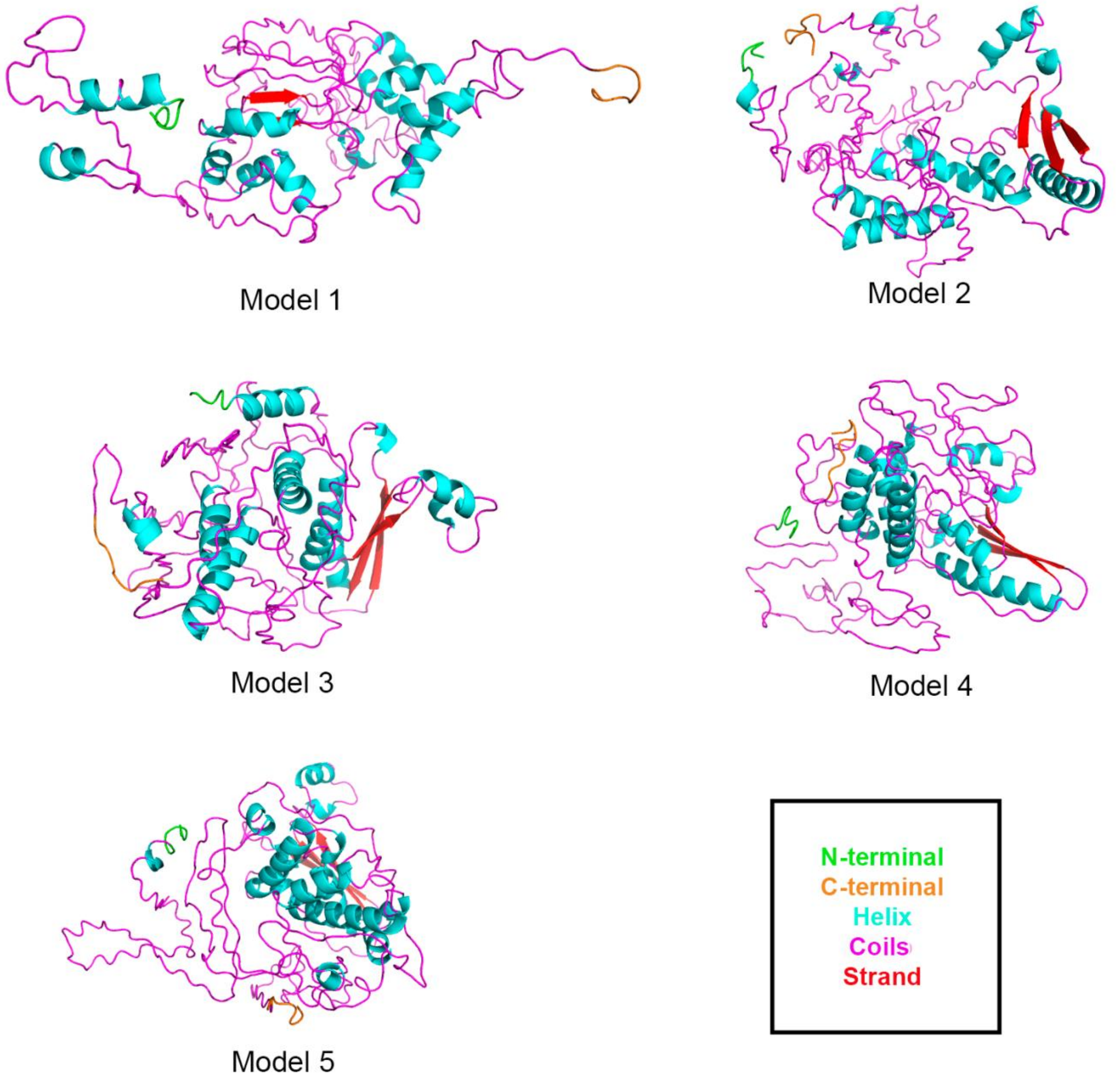


Figure 37. The five models predicted by I-TASSER using the protein threading method.

In comparison with the predicted secondary structure, the models present more alpha-helix regions, some of them of only three amino acids and some of them present at the end of the sequence. The beta-sheet regions are more similar to those predicted, except for Model 1, which presents two beta-sheets in different positions in comparison with the others. A summary of the regions where the secondary structures are found is shown in Table 6.

	PREDICTED SS	MODEL 1	MODEL 2	MODEL 3	MODEL 4	MODEL 5
Helix		8 - 12	9 – 11	6 – 17		9 – 11
		15 – 19				
		50 – 55				
		103 – 109	102 – 113	102 – 113	102 – 113	103 – 113
	100 – 133	119 – 136	120 – 136	119 – 136	119 – 136	119 – 136
		148 – 150				
			164 – 166	164 – 166	164 – 166	
						172 – 176
	182 – 191	185 – 194	182 – 191	182 – 191	182 – 191	182 – 191
			206 – 215	206 – 213	206 – 214	206 – 214
		224 – 232	221 – 224	221 – 224	221 – 224	221 – 224
	239 – 254	241 – 252	237 – 255	237 – 255	237 – 255	237 – 255
	260 – 274	268 – 271	261 – 276	259 – 276	261 – 276	260 – 281
			279 – 281			
						311 – 315
						337 – 340
		373 – 380			373 – 380	373 – 385
			395 – 397			
		397 – 407		403 – 407		
			408 – 411		407 – 417	
		419 – 426				
β-Sheet	156 – 161		154 – 161	154 – 161	154 – 161	154 – 161
	172 – 177					
	194 – 199		194 – 198	194 – 198	194 – 198	194 – 198
		200 – 202				
	229 – 232		228 – 235	228 – 235	228 – 235	228 – 235
		318 – 321				

Table 6. Summary of the positions within the Sam68 sequence where the secondary structures alpha-helix and beta-sheet were predicted by protein threading.

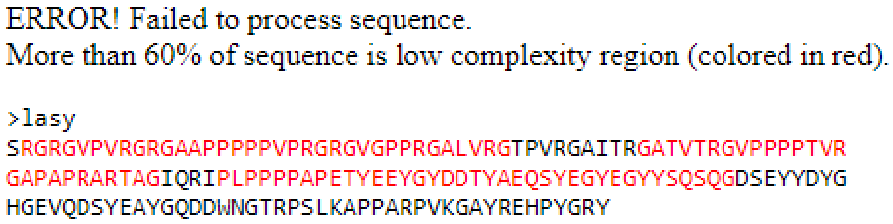
3.4. *Ab initio* predictions

3.4.1. QUARK

Although this method is not the most appropriate for Sam68 modeling (with templates with more than 30% identity, comparative protein model is the best approach), it was carried out only for comparative purposes.

Query proteins without related protein PDB structures are the most complicated cases and a successful structure prediction is limited to less than 200 amino acids; therefore the region submitted for this work was 81 – 280, which covers the area where most of the templates can be found and corresponds to the GSG domain. In addition, the

281 – 443 fragment was also submitted but this sequence could not be processed due to a low complexity region (more than 60%, colored in red as shown in Figure 38). This region corresponds to a high disordered area, as pointed out by other servers used for model prediction.



ERROR! Failed to process sequence.
More than 60% of sequence is low complexity region (colored in red).

```
>lasy
SRGRGVPVRGRGAAPPPPPVPRGRGVGPPRGALVRGTPVRGAITRGATVTRGVPPPPTVR
GAPAPRARTAGIQRIPLPPPPAPETYEEYGYDDTYAEQSYEGYEGYYSQSQGDSEYYDYG
HGEVQDSYEAYGQDDWNGTRPSLKAPPARPVKGAYREHPYGRY
```

Figure 38. Low complexity region detected in the 281 – 443 region of Sam68, making it not suitable for *ab initio* prediction by QUARK.

Figure 39 shows the five models predicted by QUARK, with the different predicted alpha-helix colored according to its order in the sequences. QUARK predicted 5 alpha-helix regions, although some of the models predicted include a small one (depicted in red). Four beta-strands are predicted in the model with high confidence scores (5 or 6 residues length), however these are not present in the models in PDB format provided by the server, and some are shown as small alpha-helix regions. The regions corresponding to the beta-strands predicted are highlighted in black. Table 7 summarizes the coordinates in the sequences where the secondary structures were found in the five models. Overall, the five models predicted similar secondary structures to those predicted by comparative modeling.

3.5. Summary

The GSG domain of the protein, which contains the KH RNA binding domain, constitutes the structural core of Sam68, being the most part of the secondary structures predicted located in its region.

It can be noted, from the schematic representation of the Sam68 protein showed in Figure 3 and the disorder prediction made, that the disordered residues are located mainly out of the GSG domain: proline-rich motifs (P0-P5), arginine/glycine/glycine (RGG) and arginine/glycine (RG) boxes, a tyrosine rich region at the C-terminal domain, and a nuclear localization signal (NLS).

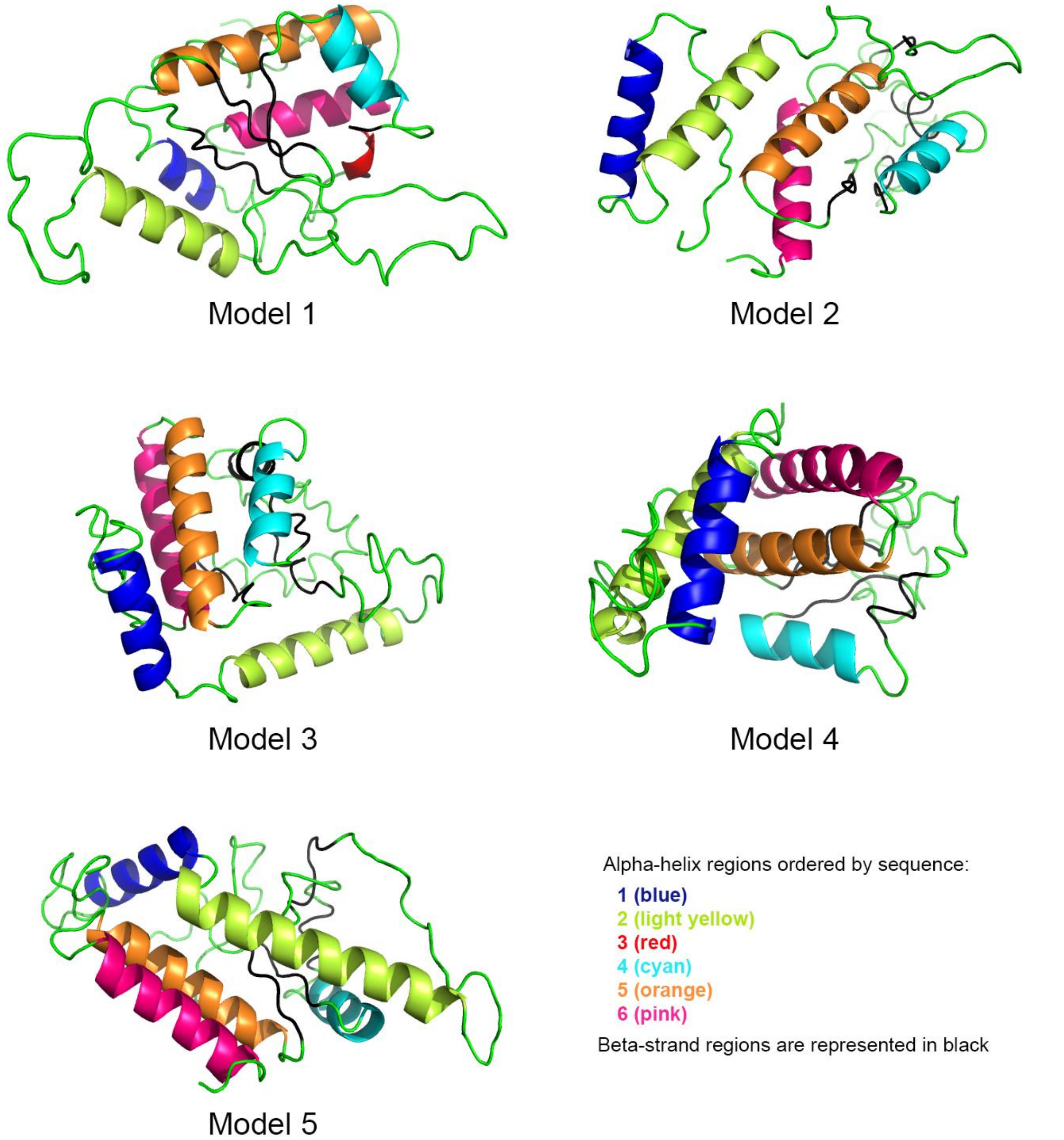


Figure 39. The five models predicted by QUARK.

	PREDICTED SS	MODEL 1	MODEL 2	MODEL 3	MODEL 4	MODEL 5
Helix	100 – 111	103 – 109	98 – 111	99 – 111	99 – 112	99 – 111
	120 – 133	118 – 132	118 – 133	118 – 134	119 – 139	115 – 137
		172 – 175		172 – 177		
	183 – 191	182 – 190	182 – 191	182 – 191	182 – 191	182 – 191
	237 – 254	237 – 254	238 – 254	237 – 254	237 – 254	237 – 254
	260 – 274	262 – 275	260 – 275	258 – 275	257 – 275	260 – 275
β-Sheet	156 – 161					
	172 – 177					
	194 – 198					
	228 – 233					

Table 7. Summary of the positions within the Sam68 sequence where the secondary structures were predicted by *ab initio* modeling.

The proline-rich regions may contribute to the lack of an ordered structure in the 3D models, as proline enables abrupt changes in the direction of the polypeptide chain. In some cases, small helix structures consisting of three amino acids were predicted, which are not alpha-helix; the helix 3_{10} could fit in these regions, although this structure is less common, tends to appear at the N- or C-terminal, and it has been described in channels and membrane proteins. Therefore, this regions predicted should not be taken into account, at least with these characteristics.

Sam68 has a remarkable portion of unstructured regions in its sequence. IntFOLD3 server represented them totally linear, whereas I-TASSER, Phyre2, and some models of BhageerathH represented the coil regions surrounding the domain and decreased the 3D space occupied by the protein. A complete linear arrangement of these regions could be ruled out, as this could affect the protein stability and it is probably energetically unfavorable.

The PrDOS server predicted a disordered region in 199 – 221 positions (KH domain), where no secondary structures would be expected. All the servers nonetheless, except for Phyre2, predicted an alpha-helix in 206 – 215. Similar results were found with the models predicted by I-TASSER, where four out of five models predicted the alpha-helix in the same region of the sequence.

The Qua1 domain has at least one predicted alpha-helix, which other servers split it into two. The KH domain accumulates the majority of the secondary structure predicted, with three or four beta-sheets and at least two alpha-helices (in some cases three or four). Two versions of the KH domain were pointed out by Grishin [58]: the KH type I (found in eukaryotic proteins) and the KH type II (found in prokaryotic proteins). KH domains usually contain a GXXG loop. Figure 40A shows the typical KH-I domain fold. The model predicted by Raptor X was used in this case to depict the order of the secondary structures found in KH-I domain. Figure 40B shows the KH domain sequence of Sam68, with the predicted secondary structures by Raptor X, and the 3D-structure predicted by this server is shown in Figure 40C with the typical secondary structures found in a KH type I domain.

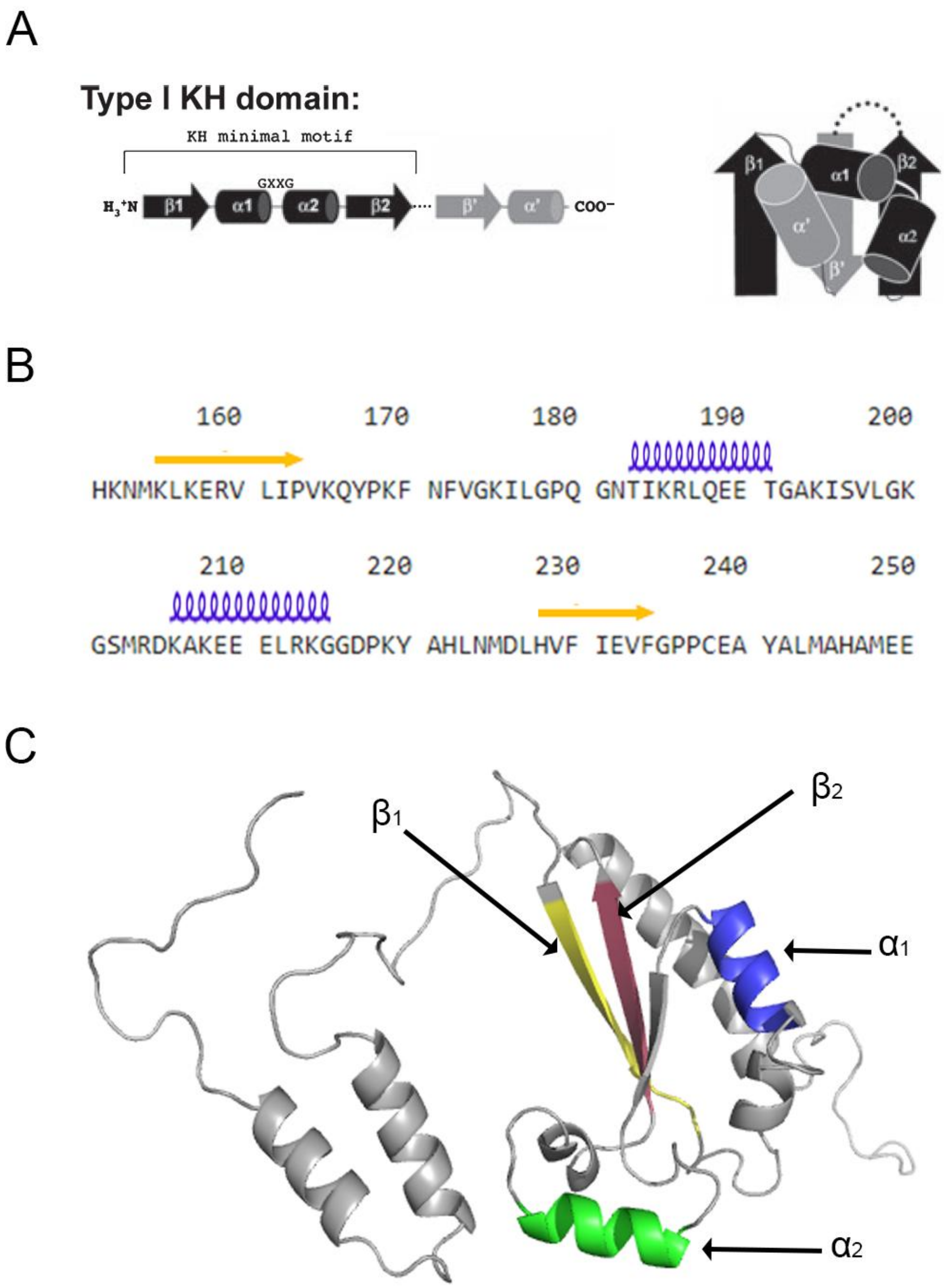

A
Type I KH domain:
KH minimal motif
GXXG
H3+N
β1
α1
α2
β2
β'
α'
COO−
B
160 170 180 190 200
HKNMKLKERV LIPVKQYPKF NFVGKILGPQ GNTIKRLQEE TGAKISVLGK
210 220 230 240 250
GSMRDKAKEE ELRKGGDPKY AHLNMDLHVF IEVFGPPCEA YALMAHAMEE
C
β1
β2
α1
α2

Figure 40. (A) Scheme of the typical KH-I domain (taken from Valverde et al. [59]). (B) Sam68 sequence with the secondary structures predicted by Raptor X within the KH domain. (C) Model with the βααβ core highlighted.

Overall, the models predicted tend to keep this minimal KH motif (the βααβ core), although most of the cases a beta-sheet is predicted between the two alpha-helix motifs, and a GXXG motif within the sequence is not found to be exactly a loop between the alpha-helices. It is noteworthy that none of the PDB structures provided by Feracci *et al*. [22] had this canonical structure, although the groove formed by these regions is the surface for RNA binding modeled in that work.

Figure 41 shows a summary of the predicted 3D structure of Sam68 in relation with its schematic structure shown in Figure 3. Red lines represent disordered/unstructured regions and blue lines indicate ordered regions.

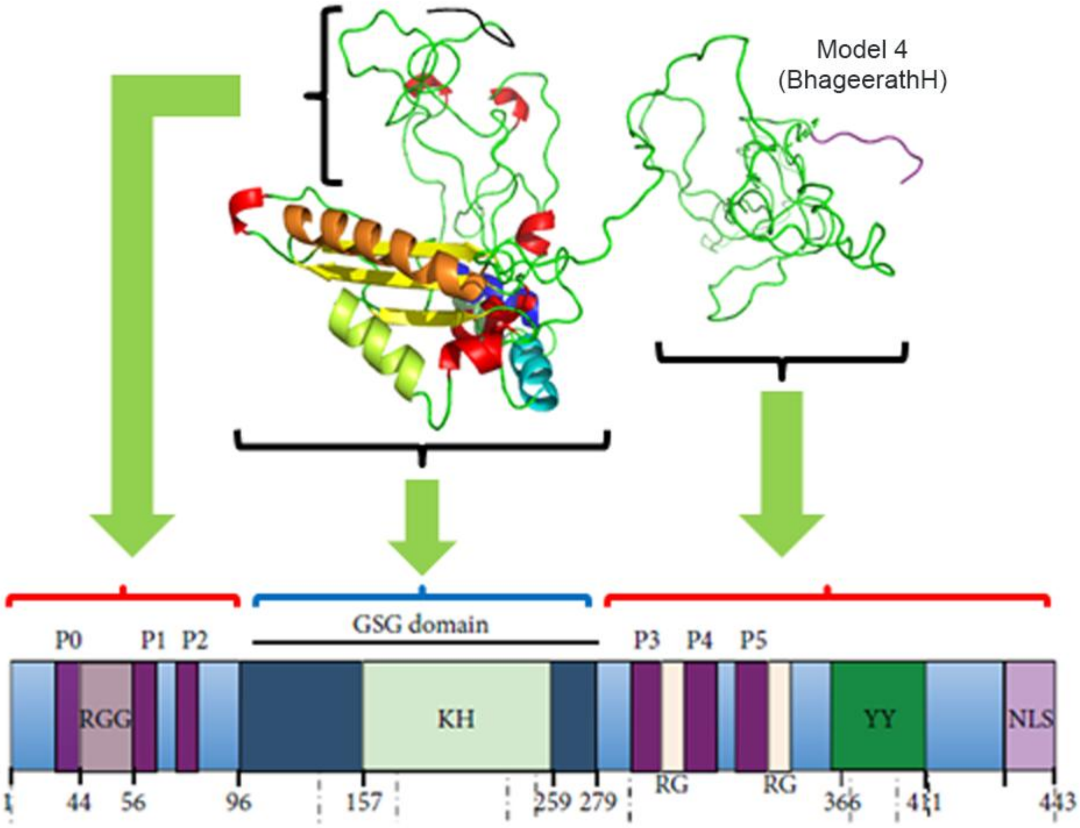


Figure 41. One of the predicted 3D models of Sam68 and its correspondence with its linear sequence, where the main regions can be found.

4. Sam68 – RNA interactions

RNA-binding proteins regulate several cellular processes through their binding to mRNA or non-coding RNA. The RBP Sam68 is involved in the SMA disease through its regulation of *SMN2* expression. Therefore, the study of the RNA-protein interactions is of high interest given its recent involvement in the disease. Pedrotti *et al.* [16] previously described that the V229F mutation in the KH domain and mutations in the NLS region in Sam68 inhibited the exon-7 skipping mediated by this protein.

Using PDBsum, the main RNA contact sites in the PDB structures provided by Feracci *et al.* [22] using the protein T-STAR are listed in Table 8. The corresponding amino acid in the Sam68 sequence is also provided after the local alignment of both sequences. These residues will be taken into account in the predictions.

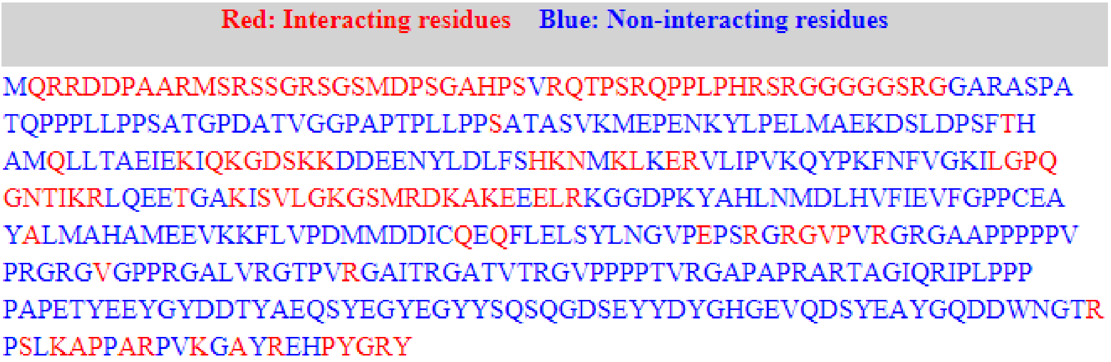
T-STAR	Sam68
Gly78	Gly178
Arg80	Gln180
Lys85	Lys185
Gly101	Gly201
Arg104	Arg204

Table 8. Main RNA-binding residues found in T-STAR and their correspondence in Sam68 sequence.

Some of the computational tools currently available for studying RNA-protein interactions were consulted in the reviews of Muppirala et al. [60] and Puton et al. [61].

4.1. Predictions of RNA-binding residues from protein sequence

The PPRint (Prediction of Protein Interaction, http://webs.iiitd.edu.in/raghava/pprint/index.html) server was used to predict the interacting residues in Sam68. This method uses support vector machines (SVM). According to the results shown in Figure 42, first 50 residues would be RNA-interacting; some residues within the Qua1 domain, residues from the KH domain, and the NLS region would also be RNA-interacting amino acids. Residues from Table8 are all predicted as RNA-binding amino acids.

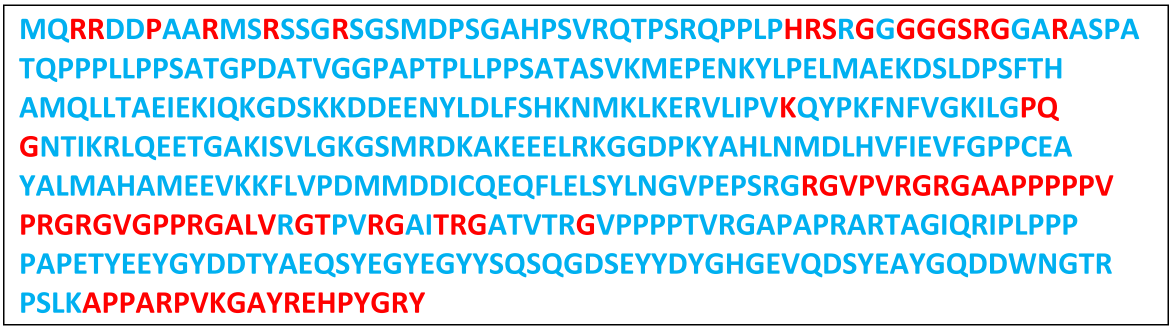


Red: Interacting residues   Blue: Non-interacting residues

```
MQRRDDPAARMSRSSGRSGSMDPSGAHPSVRQTPSRQPPLPHRSRGGGGGSRGGARASPA
TQPPPLLPPSATGPDATVGGPAPTPLLPPSATASVKMEPENKYLPELMAEKDSLDPSFTH
AMQLLTAEIEKIQKGDSKKDDEENYLDLFSHKNMKLKERVLIPVKQYPKFNFVGKILGPQ
GNTIKRLQEETGAKISVLGKGSMRDKAKEEELRKGGDPKYAHLNMDLHVFIEVFGPPCEA
YALMAHAMEEVKKFLVPDMMDDICQEQFLELSYLNGVPEPSRGRGVPVRGRGAAPPPPPV
PRGRGVGPPRGALVRGTPVRGAITRGATVTRGVPPPPTVRGAPAPRARTAGIQRIPLPPP
PAPETYEEYGYDDTYAEQSYEGYEGYYSQSQGDSEYYDYGHGEVQDSYEAYGQDDWNGTR
PSLKAPPARPVKGAYREHPYGRY
```

Figure 42. Prediction of RNA-interacting residues of Sam68 protein using PPRInt.

RNABindRPlus also predicts RNA-binding residues using only the protein sequence, but using Naïve Bayes classifier. Similar to PPRint, scores lower than 0.5 are indicated as non RNA-binding residues, whereas scores equal or greater than 0.5 indicate RNA-binding amino acids. It predicted 78 binding residues, in contrast with the 115 predicted by SVM. Figure 43 shows the output using this tool, with RNA-interacting residues highlighted in red and non-interacting residues highlighted in blue, in order to compare these results with those obtained using PPRInt.



MQRRDDPAARMSRSSGRSGSMDPSGAHPSVRQTPSRQPPLPHRSRGGGGGSRGGARASPA
TQPPPLLPPSATGPDATVGGPAPTPLLPPSATASVKMEPENKYLPELMAEKDSLDPSFTH
AMQLLTAEIEKIQKGDSKKDDEENYLDLFSHKNMKLKERVLIPVKQYPKFNFVGKILGPQ
GNTIKRLQEETGAKISVLGKGSMRDKAKEEELRKGGDPKYAHLNMDLHVFIEVFGPPCEA
YALMAHAMEEVKKFLVPDMMDDICQEQFLELSYLNGVPEPSRGRGVPVRGRGAAPPPPPV
PRGRGVGPPRGALVRGTPVRGAITRGATVTRGVPPPPTVRGAPAPRARTAGIQRIPLPPP
PAPETYEEYGYDDTYAEQSYEGYEGYYSQSQGDSEYYDYGHGEVQDSYEAYGQDDWNGTR
PSLKAPPARPVKGAYREHPYGRY

Figure 43. Prediction of RNA-interacting residues of Sam68 protein using RNABindRPlus.

The C-terminal region contains RNA-interacting residues, similar to the PPRInt prediction. Nevertheless, the main area with RNA-binding amino acids is 284 – 314, which is out of the KH domain. Overall, the output of these tools significantly differ. In addition, only one of the amino acids from Table 8 (Gln180) was predicted as an RNA-interacting residue.

4.2. Predictions of RNA-binding residues from protein and RNA sequences

The catRAPID server provides several modules to study RNA-protein interactions; one of its advantages over the previous methods is that it predicts the binding sites in both RNA and protein sequences. For proteins less than 750 amino acids and RNA sequences up to 1200 nucleotides, the tools catRAPID *graphic* and catRAPID *strength* are recommended, whereas larger proteins or RNAs require the module catRAPID *fragment* [62]. Sam68 is less than 750 amino acids in length, but the transcripts found for the *SMN2* gene are larger than 1200 nucleotides. There are four known transcripts of the *SMN2* gene (the complete sequences can be found in the "Appendices" section):

NCBI RefSeq: NM_017411 → Isoform d; 1634 bp

NCBI RefSeq: NM_022875.2 → Isoform a; 1580 bp (this is thought to be the predominant transcript)

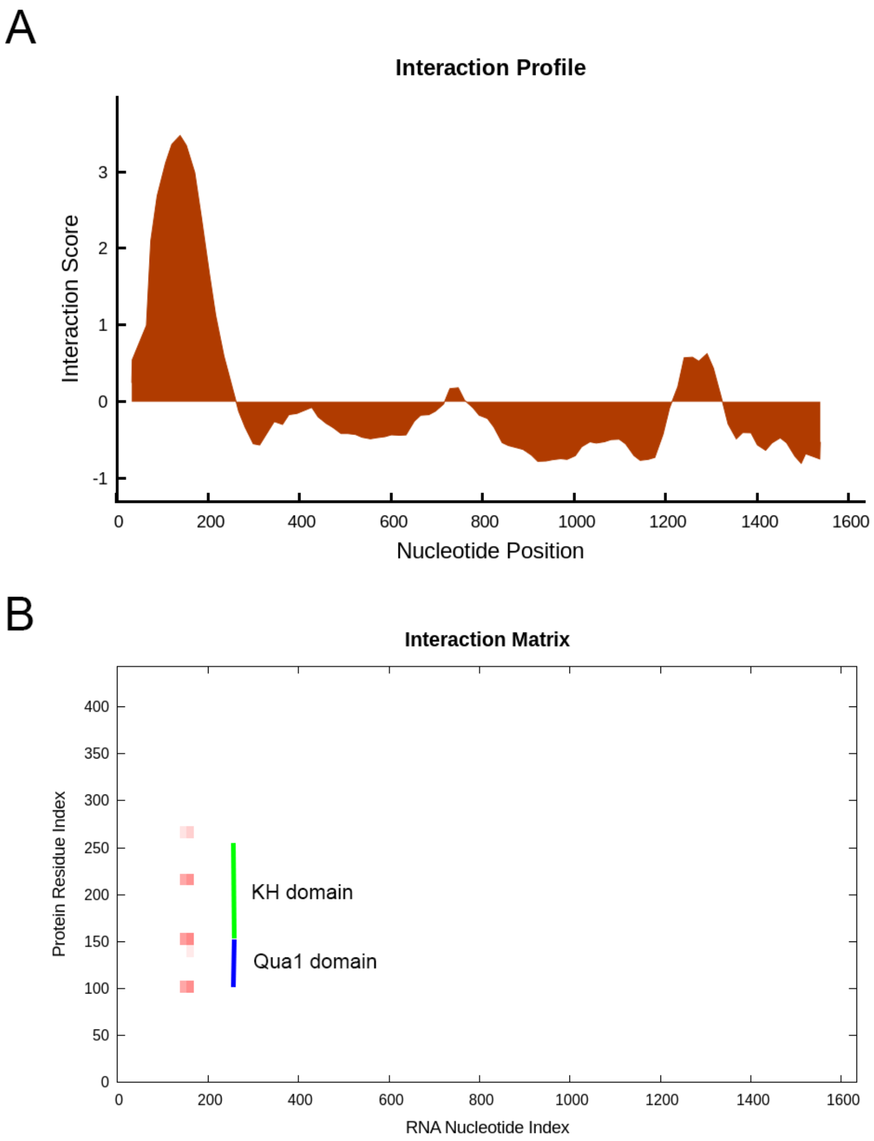
NCBI RefSeq: NM_022877.2 → Isoform c; 1484 bp

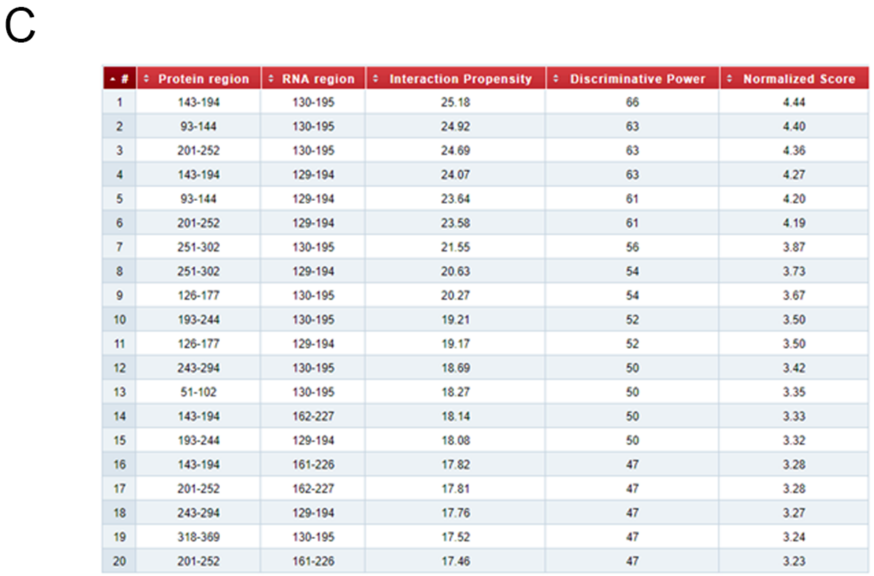
NCBI RefSeq: NM_022876 → Isoform b; 1538 bp

The isoform d is the longest transcript, thus encoding the longest isoform. This module divides both the protein and the RNA sequence and predicts their interaction propensities. The output is shown in Figure 44. The interaction profile represents the interaction score along the RNA sequence (Figure 44A). Figure 44B represents the plot which shows, as a heat-map, the amino acids and nucleotides predicted to interact. Last, Figure 44C is the table summarizing the top 20 interactions. The highest discriminative power scores (measurement of the interaction propensity of a protein-RNA pair) were obtained in the protein region containing the Qua1 and KH domains interacting with the RNA region 130 – 195.

4.3. Prediction of RNA-binding sites from protein structures

The KYG method was the only one available online to perform a prediction of RNA-binding sites from a structure. The model predicted by Phyre2 (Figure 45A) and one of the models predicted by I-TASSER (Figure 45B) were used. N-terminal and C-terminal regions obtained good scores in both cases. The regions following, approximately, the KH domain also obtained good scores in the two predictions; only few residues within this domain had scores greater than 0.08 and were considered RNA-binding sites.





#	Protein region	RNA region	Interaction Propensity	Discriminative Power	Normalized Score
1	143-194	130-195	25.18	66	4.44
2	93-144	130-195	24.92	63	4.40
3	201-252	130-195	24.69	63	4.36
4	143-194	129-194	24.07	63	4.27
5	93-144	129-194	23.64	61	4.20
6	201-252	129-194	23.58	61	4.19
7	251-302	130-195	21.55	56	3.87
8	251-302	129-194	20.63	54	3.73
9	126-177	130-195	20.27	54	3.67
10	193-244	130-195	19.21	52	3.50
11	126-177	129-194	19.17	52	3.50
12	243-294	130-195	18.69	50	3.42
13	51-102	130-195	18.27	50	3.35
14	143-194	162-227	18.14	50	3.33
15	193-244	129-194	18.08	50	3.32
16	143-194	161-226	17.82	47	3.28
17	201-252	162-227	17.81	47	3.28
18	243-294	129-194	17.76	47	3.27
19	318-369	130-195	17.52	47	3.24
20	201-252	161-226	17.46	47	3.23

Figure 44. (A) RNA-interaction profiles for Sam68 association with *SMN2*. (B) Interaction map of Sam68 with *SMN2* mRNA, with the domains of Sam68 indicated. (C) Table with the top 20 interactions predicted.

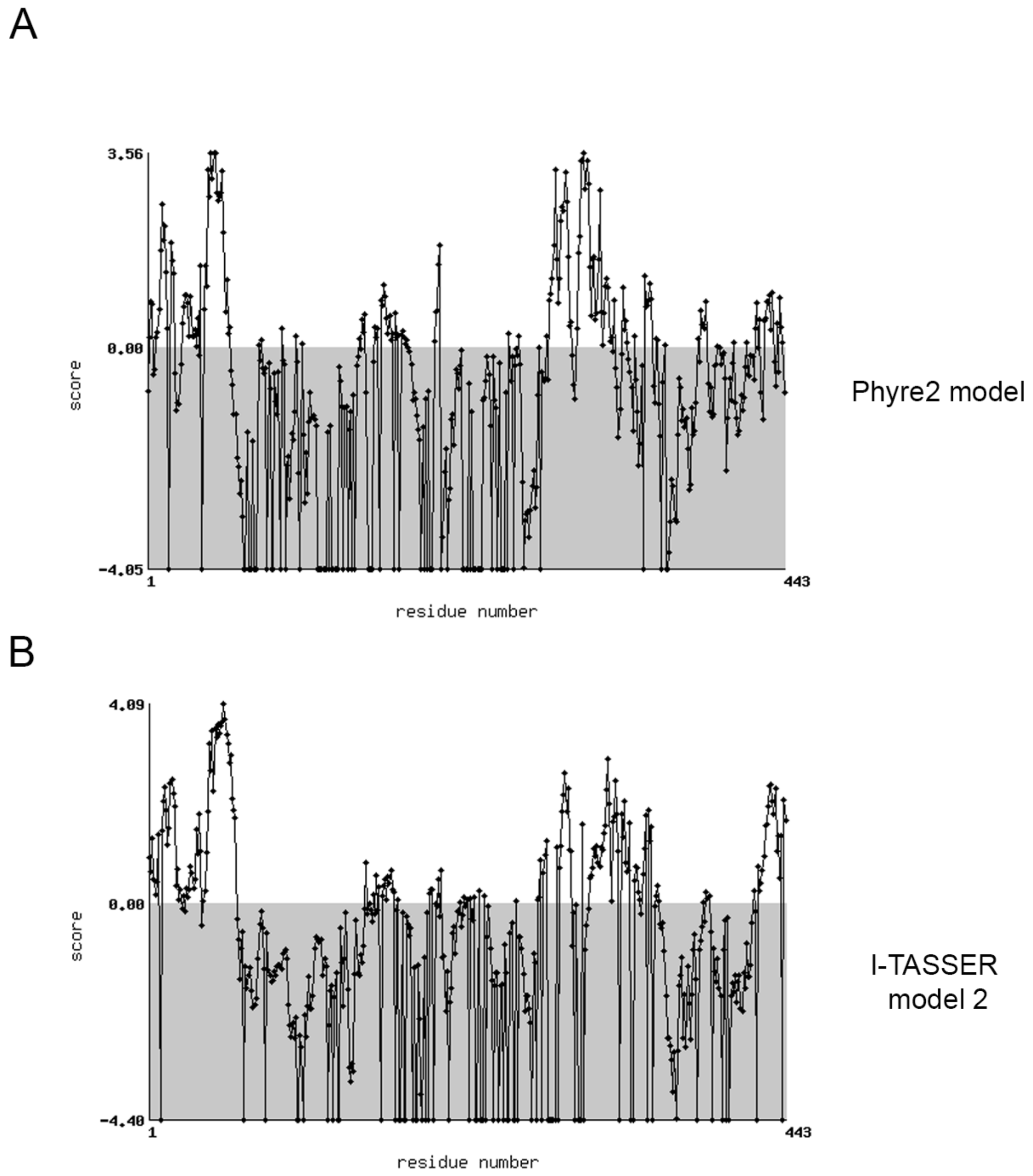


Figure 45. Plot of the RNA interface residue predictions for (A) the Phyre 2 model and (B) the model 2 from I-TASSER predictions.

In fact, the greatest scores within the RNA-binding domain were obtained in similar regions for both predictions; in the Phyre2 model, it was the 163 – 173 region which is located between two beta-sheets according to its secondary structure prediction; in the I-TASSER model, it was the 161 – 170 region (together with the Lys152 residue), which is located right after a beta-sheet and before an alpha-helix. Figure 46 shows the KH domain in both models with their corresponding RNA-interacting residues highlighted in different colors.

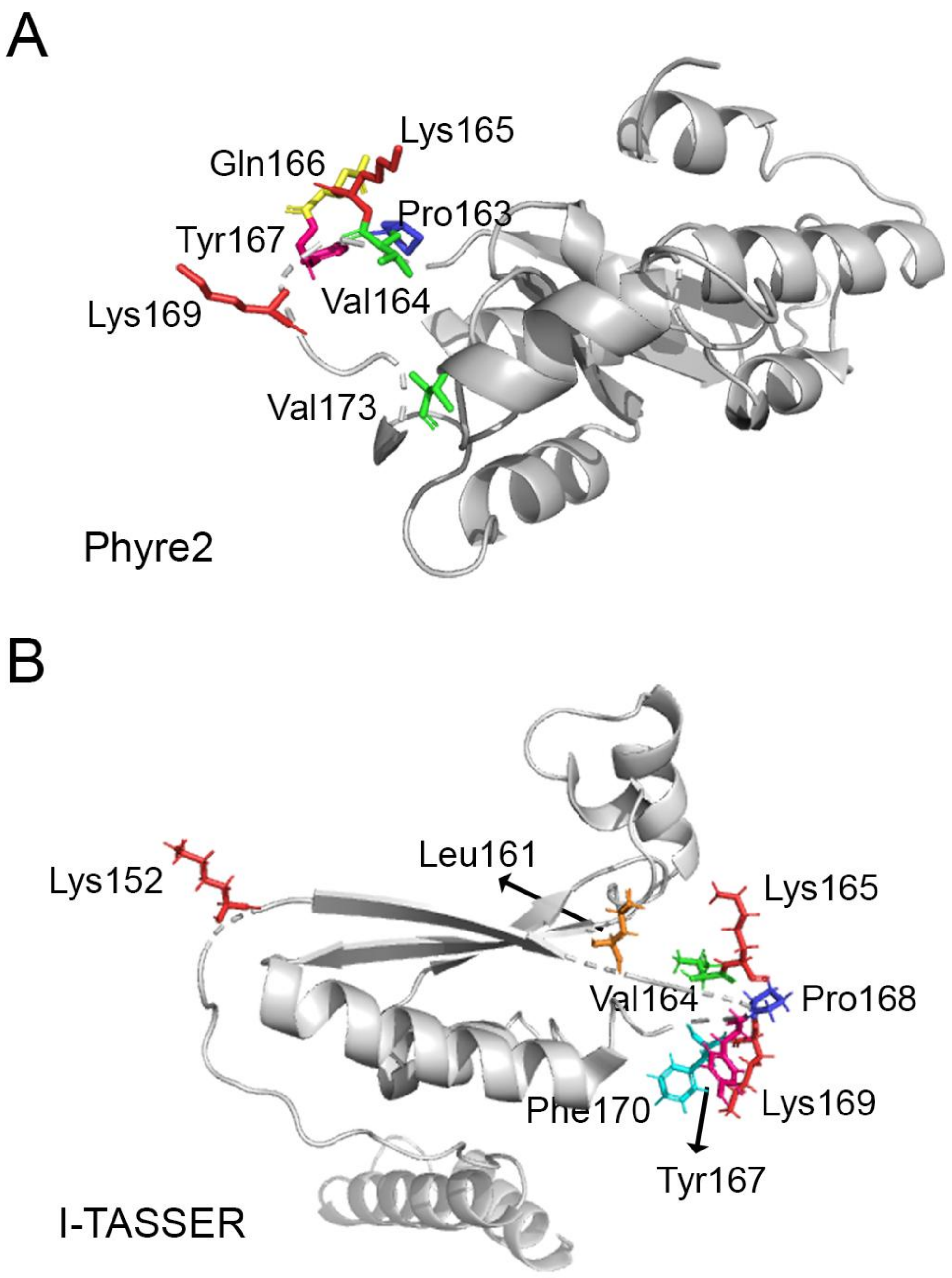


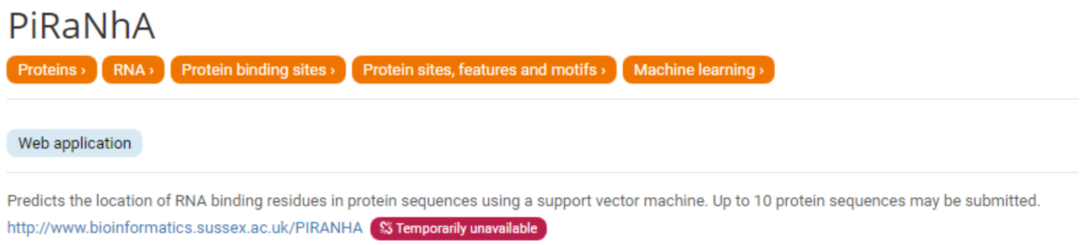
Figure 46. RNA-binding residues predicted by KYG in (A) Phyre2 model and (B) I-TASSER model. Lys residues are colored in red; Val in green; Pro in blue; Gln in yellow, Tyr in pink; Leu in orange; Phe in cyan.

The prediction with Phyre2 model obtained good scores for the RNA-binding sites (except for Lys185) from Table 8, although not the greatest ones. In contrast, only Arg204 was predicted as interacting residue using the I-TASSER model.

4.4. Limitations of the study of RNA – protein interactions

Other tools were consulted as well, which were unavailable at the moment the study was conducted:

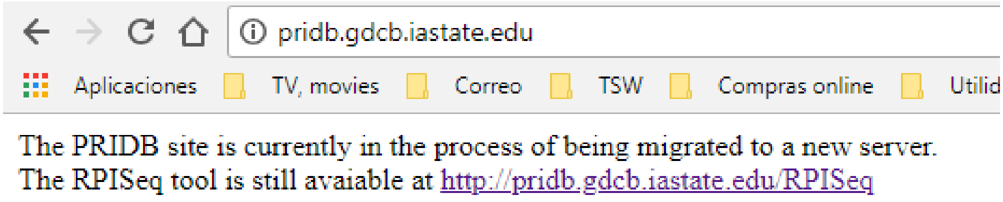
-PiRaNhA: for prediction RNA-binding residues in a protein sequence. It is temporarily unavailable (Figure 47).



Predicts the location of RNA binding residues in protein sequences using a support vector machine. Up to 10 protein sequences may be submitted.
http://www.bioinformatics.sussex.ac.uk/PIRANHA

Figure 47.

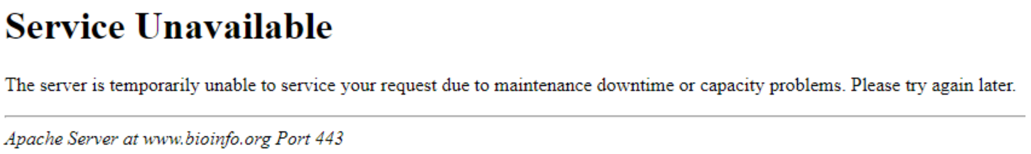
-The Protein Interface Database (PRIDB) contains an exhaustive protein-RNA interaction database extracted from PDB structures. As shown in Figure 48, PRIDB is not accessible due to a migration to a new server.



The PRIDB site is currently in the process of being migrated to a new server.
The RPISeq tool is still avaiable at http://pridb.gdcb.iastate.edu/RPISeq

Figure 48.

-NPInter is also a database for RNA-protein interactions, which was unavailable as well (Figure 49).



Service Unavailable

The server is temporarily unable to service your request due to maintenance downtime or capacity problems. Please try again later.

Apache Server at www.bioinfo.org Port 443

Figure 49.

5. Conclusions

In this work, a comprehensive study about the current knowledge about Sam68 structure has been carried out. A review of the literature about this protein and its involvement in spinal muscular atrophy has been conducted, and a number of computational tools have been employed in order to study the structure of this protein and its interactions with RNA.

Sam68 is a ribonucleoprotein which regulates alternative splicing of several genes. Impairment of Sam68 activity may rescue the activity of the SMN protein, which is affected in spinal muscular atrophy disease.

The involvement of this protein rests on the fact that the exclusion of exon 7 of the *SMN2* gene leads to an unstable protein, which cannot compensate the loss of SMN activity in SMA. Thus, the interaction between Sam68 and RNA has a potential role in the disease. Gaining insight into the 3D structure of Sam68 may contribute to the development of future short interfering peptides / molecules to improve the motor function.

The models predicted by several servers obtained good scores in the region of the protein corresponding to the KH-domain, which is the area of interest (RNA-binding domain). There were slight differences in the predicted secondary structures between the different tools used for structure prediction (by homology modeling and protein threading). Despite this, there is a significant degree of consensus regarding the prediction of secondary structures. The KH domain structure could be found in most of the predicted models, although not in its canonical structure described in the literature. Therefore, although several homologous proteins with good similarity percentages were available as templates for homology modeling, further studies are required.

The most difficult areas within Sam68 sequence to model were the N- and C-terminal, since they have been predicted as disordered and low complexity regions. Nevertheless, these areas are of great interest due to their length (approximately 100 – 150 residues) and it is necessary to elucidate / predict their folding, together with the KH domain. Unfortunately, this problem is beyond the scope of this work.

Different tools were also employed for prediction of RNA-protein interactions. The predictions based on the protein sequence significantly differ, whereas predictions based on both protein and RNA sequences achieved better scores and pointed out to the Qua1 and KH domain as the areas of interaction with RNA. The study of RNA-protein interactions had several limitations, due to the fact that some of the online tools were not available at the time this work was in progress and to the lack of enough experimental data. Nevertheless, the predictions made by KYG method, which is based on the protein structure, pointed to ~160 – 173 region within the KH domain as a likely area with RNA-interacting residues. These results, together with the protein-RNA contact sites consulted in PDBsum for the most recent PDB structures studying Sam68 and T-STAR proteins represent a good starting point.

Although the current tools available for the study of the 3D structure of a protein are powerful and efficient, experimental methods, such as X-ray crystallography, electron microscopy, or nuclear magnetic resonance, would be desirable. Although these approaches are costly and time consuming, the relevance of Sam68 in SMA disease is remarkable enough to consider it as a research proposal.

6. Glossary

Alternative splicing: regulated process during gene expression that results in a single gene coding for different proteins.

Amino acids / residues: biomolecules consisting of an amino group, a carboxylic group, and an organic group or side chain (usually named R) which is specific for each amino acid.

Domain: any segment of a protein which folds into a structure unit, usually having a specific function.

Electron density maps: these maps are produced as a result of X-ray crystallographic experiments.

Fo-Fc / 2Fo-Fc: types of electron density maps, with the structure factor amplitudes calculated from the model (Fc), and the measured structure factors from the diffraction patterns (Fo).

Knock-down: a technique by which expression of one or more genes is reduced.

PDB: Protein Data Bank

Primary structure: the amino acid sequence of a protein

Protein modeling / prediction: the process by which the three-dimensional structure of a protein can be inferred.

RBP: RNA-binding protein

RNA: ribonucleic acid

RNAi: RNA interference; a process in which RNA inhibits gene expression or translation.

Secondary structure: local structures stabilized mainly by hydrogen bonds; the most common are alpha-helix, beta-sheet and turns.

shRNA: short/small hairpin RNA is an artificial RNA used to silence target genes,

SMA: Spinal Muscular Atrophy

SMN: survival motor neuron

Template: a related homologous protein which is used for structure prediction of the protein target.

7. References

1. Finkel R, Bertini E, Muntoni F, Mercuri E. 209th ENMC international workshop: outcome measures and clinical trial readiness in spinal muscular atrophy 7-9 November 2014, Heemskerk, The Netherlands. Neuromuscul Disord. 593–602. 25, 2015.
2. Verhaart IEC, Robertson A, Wilson IJ, Aartsma-Rus A, Cameron S, Jones CC, Cook SF, Lochmüller H. Prevalence, incidence and carrier frequency of 5q–linked spinal muscular atrophy – a literature review. Orphanet J Rare Dis. 124. 12, 2017
3. Lefebvre S, Burglen L, Reboullet S, Clermont O, Burlet P, Viollet L, Benichou B, Cruaud C, Millasseau P, Zeviani M, *et al.* Identification and characterization of a spinal muscular atrophy-determining gene. Cell. 155–165. 80, 1995.
4. Burglen L, Lefebvre S, Clermont O, Burlet P, Viollet L, *et al.* Structure and organization of the human survival motor neurone (SMN) gene. Genomics. 479–482. 32, 1996.
5. Lukong KE, Richard S.Sam68, the KH domain-containing superSTAR. Biochim Biophys Acta. 73–86. 1653, 2003.

6. Chen T, Damaj BB, Herrera C, Lasko P, Richard S. Self-association of the single-KH-domain family members Sam68, GRP33, GLD-1, and Qk1: Role of the KH domain. Mol Cell Biol. 5707–5718. 17, 1997.

7. Frisone P, Pradella D, Di Matteo A, Belloni E, Ghigna C, Paronetto MP. SAM68: Signal transduction and RNA metabolism in human cancer. Biomed Res Int. 528954. 2015, 2015.

8. Taylor SJ, Resnick RJ, Shalloway D. Sam68 exerts separable effects on cell cycle progression and apoptosis. BMC Cell Biol. 5–16. 5, 2004.

9. Babic I, Cherry E, Fujita DJ. SUMO modification of Sam68 enhances its ability to repress cyclin D1 expression and inhibits its ability to induce apoptosis. Oncogene. 4955–4964. 25, 2006.

10. Paronetto MP, Achsel T, Massiello A, Chalfant CE, Sette C. The RNA-binding protein Sam68 modulates the alternative splicing of Bcl-x. JCB. 929. 176, 2007.

11. Busa R, Paronetto MP, Farini D, Pierantozzi E, Botti F, Angelini DF, Attisani F, Vespasiani G, Sette C. The RNA-binding protein Sam68 contributes to proliferation and survival of human prostate cancer cells. Oncogene. 4372–4382. 26, 2007.

12. Zhang Z, Li J, Zheng H, Yu C, Chen J, Liu Z, Li M, Zeng M, Zhou F, Song L. Expression and cytoplasmic localization of SAM68 is a significant and independent prognostic marker for renal cell carcinoma. Cancer Epidemiol Biomark Prev. 2685–2693. 18, 2009.

13. Li Z, Yu CP, Zhong Y, Liu TJ, Huang QD, Zhao XH, Huang H, Tu H, Jiang S, Zhang Y, *et al.* Sam68 expression and cytoplasmic localization is correlated with lymph node metastasis as well as prognosis in patients with early-stage cervical cancer. Ann Oncol. 638–646. 23, 2012.

14. Chen SW, Zhang Q, Yang AK, Li Z, Zhong Y, Li H, Zeng Y, Zhuang SM, Wang LP, Song LB, *et al.* Overexpression and cytoplasmic localization of Sam68 correlate with tumour progression and poor prognosis in patients with clinically N0 oral tongue cancer. Head Neck Oncol. 61. 4, 2012.

15. Liao WT, Liu JL, Wang ZG, Cui YM, Shi L, Li TT, Zhao XH, Chen XT, Ding YQ, Song LB. High expression level and nuclear localization of Sam68 are associated with progression and poor prognosis in colorectal cancer. BMC Gastroenterol. 126. 13, 2013.

16. Pedrotti S, Bielli P, Paronetto MP, Ciccosanti F, Fimia GM, Stamm S, Manley JL, Sette C. The splicing regulator Sam68 binds to a novel exonic splicing silencer and functions in SMN2 alternative splicing in spinal muscular atrophy. EMBO J. 1235-1247. 29, 2010.

17. Naryshkin NA, Weetall M, Dakka A, Narasimhan J, Zhao X, Feng Z, Ling KK, Karp GM, Qi H, Woll MG, Chen G, Zhang N, Gabbeta V, *et al.* Motor neuron disease. SMN2 splicing modifiers improve motor function and longevity in mice with spinal muscular atrophy. Science. 688-693. 345, 2014.

18. Taylor SJ, Shalloway D. An RNA-binding protein associated with src through its SH2 and SH3 domains in mitosis. Nature. 867 – 871. 368, 1994.
19. Lin Q, Taylor SJ, Shalloway D. Specificity and determinants of Sam68 RNA binding. J Biol Chem. 27274 – 27280. 272, 1997.
20. Galarneau A, Richard S. The STAR RNA binding proteins GLD-1, QKI, SAM68 and SLM-2 bind bipartite RNA motifs. BMC Mol Biol. 47. 10, 2009.
21. Tichon A, Perry RB,Stojic L, Ulitsky I. SAM68 is required for regulation of Pumilio by the NORAD long noncoding RNA. Genes Dev. 70 – 78. 32, 2018.
22. Feracci M, Foot JN, Grellscheid SN, Danilenko M, Stehle R, Gonchar O, Kang H, Dalgliesh C, Meyer NH, Liu Y, Lahat A, Sattler M, Eperon IC, Elliott DJ, Dominguez C. Structural basis of RNA recognition and dimerization by the STAR proteins T-STAR and Sam68. Nat. Commun. 10355. 7, 2016.
23. Li DK, Tisdale S, Lotti F, Pellizzoni L. SMN control of RNP assembly: from post-transcriptional gene regulation to motor neuron disease. Semin Cell Dev Biol. 22-29. 32, 2014.
24. Kong L, Wang X, Choe DW, Polley M, Burnett BG, Bosch-Marcé M, Griffin JW, Rich MM, Sumner CJ. Impaired synaptic vesicle release and immaturity of neuromuscular junctions in spinal muscular atrophy mice. J Neurosci. 842-851. 29, 2009.
25. Donlin-Asp PG, Bassell GJ, Rossoll W. A role for the survival of motor neuron protein in mRNP assembly and transport. Curr Opin Neurobiol. 53-61. 39, 2016.
26. Hensel N, Stockbrügger I, Rademacher S, Broughton N, Brinkmann H, Grothe C, Claus P. Bilateral crosstalk of rho- and extracellularsignal- regulated-kinase (ERK) pathways is confined to an unidirectional mode in spinal muscular atrophy (SMA). Cell Signal. 540-548. 26, 2014.
27. Hensel N, Claus P. The actin cytoskeleton in SMA and ALS: how does it contribute to motoneuron degeneration? Neuroscientist. 54-72. 24, 2018.
28. Bowerman M, Becker CG, Yáñez-Muñoz RJ, Ning K, Wood MJA, Gillingwater TH, Talbot K, and The UK SMA Research Consortium. Therapeutic strategies for spinal muscular atrophy: SMN and beyond. Dis Model Mech. 943-954. 10, 2017.
29. Spitali P, Aartsma-Rus A. Splice modulating therapies for human disease. Cell. 1085-1088. 148, 2012.

30. Singh NK, Singh NN, Androphy EJ, Singh RN. Splicing of a critical exon of human Survival Motor Neuron is regulated by a unique silencer element located in the last intron. Mol Cell Biol. 1333-1346. 26, 2006.
31. Singh, N. N., Howell, M. D., Androphy, E. J. and Singh, R. N. (2017). How the discovery of ISS-N1 led to the first medical therapy for spinal muscular atrophy. Gene Ther. 520-526. 24, 2017.
32. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The protein data bank: A computer based archival file for macromolecular structures. J Mol Biol. 535–542. 112, 1977
33. Laskowski RA, Hutchinson EG, Michie AD, Wallace AC, Jones ML, Thornton JM. PDBsum: A webbased database of summaries and analyses of all PDB structures. Trends Biochem Sci. 488–490. 22, 1997
34. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY, Dosztányi Z, *et al*. InterPro in 2017 — beyond protein family and domain annotations. Nucleic Acids Res. D190 - D199. 45, 2017.
35. Drozdetskiy A, Cole C, Procter J, Barton GJ. JPred4: a protein secondary structure prediction server. Nucleic Acids Res. W389-394. 43, 2015.
36. Ishida T, Kinoshita K. PrDOS: prediction of disordered protein regions from amino acid sequence. Nucleic Acids Res. W460-464. 35, 2007.
37. Petersen TN, Søren Brunak, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods. 785-786. 8, 2011
38. Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V. A completely reimplemented MPI Bioinformatics toolkit with a new HHpred server at its core. J Mol Biol. 30587-30589. 17, 2017.
39. Söding J. Protein homology detection by HMM-HMM comparison. Bioinformatics. 951-960. 21, 2005.
40. Källberg M, Wang H, Wang S, Peng J, Wang Z, Lu H, Xu J. Template-based protein structure modeling using the RaptorX web server. Nature Protoc. 1511-1522. 7, 2012.

41. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The Phyre2 web portal for protein modeling, prediction and analysis. Nature Protoc. 845-858. 10, 2015.
42. McGuffin LJ, Atkins J, Salehe BR, Shuid AN, Roche DB. IntFOLD: an integrated server for modelling protein structures and functions from amino acid sequences. Nucleic Acids Res. W169-173. 43, 2015.
43. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L, Schwede T. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res. W252-W258. 42, 2014.
44. Yang Y, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: Protein structure and function prediction. Nature Methods. 7-8. 12, 2015.
45. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. 40. 9, 2008.
46. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nature Protoc. 725-738. 5, 2010.
47. Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins. 1715-1735. 80, 2012.
48. Xu D, Zhang Y. Toward optimal fragment generations for ab initio protein structure assembly. Proteins. 229-239. 81, 2013.
49. Kumar M, Gromiha MM, Raghava GPS. Prediction of RNA binding sites in a protein using SVM and PSSM profile. Proteins. 189-194 71, 2008.
50. Bellucci M, Agostini F, Masin M, Tartaglia GG. Predicting protein associations with long noncoding RNAs. Nat Methods. 444-445. 8, 2011.
51. Kim OTP, Yura K, Go N. Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction. Nuc Acids Res. 6450-6460. 34, 2006.
52. Emsley P, Lohkamp B, Scott W, Cowtan K. Features and Development of Coot. Acta Cryst Section D-Biological Crystallography. 486-501. 66, 2010.
53. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.
54. Kleywegt GJ, Harris MR, Zou JY, Taylor TC, Wählby A, Jones TA. The Uppsala Electron-Density Server. Acta Cryst Section D-Biol Crystal. 2240-2249. 60, 2004.

55. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 2577-2637. 22, 1983.
56. Morishita EC, Murayama K, Kato-Murayama M, Ishizuka-Katsura Y, Tomabechi Y, Hayashi T, Terada T, Handa N, Shirouzu M, Akiyama T, Yokoyama S. Crystal structures of the armadillo repeat domain of adenomatous polyposis coli and its complex with the tyrosine-rich domain of Sam68. Structure. 1496-508. 19, 2011.
57. Meyer NH, Tripsianes K, Vincendeau M, Madl T, Kateb F, Brack-Werner R, Sattler M. Structural basis for homodimerization of the Src-associated during mitosis, 68-kDa protein (Sam68) Qua1 domain. J Biol Chem. 28893-28901. 285, 2010.
58. Grishin NV. KH domain: one motif, two folds. Nucleic Acids Res. 638–643. 29, 2001
59. Valverde R, Edwards L, Regan L. Structure and function of KH domains. FEBS Journal. 2712–2726. 275, 2008.
60. Muppirala UK, Lewis BA, Dobbs D. Computational tools for investigating RNA-Protein interaction partners. J Comput Sci Syst Biol. 182-187. 6, 2013.
61. Puton T, Kozlowski L, Tuszynska I, Rother K, Bujnicki JM. Computational methods for prediction of protein-RNA interactions.J Struct Biol.261-268. 179, 2012.
62. Cirillo D, Livi CM, Agostini F, Tartaglia GG. Discovery of protein–RNA networks. Mol. BioSyst. 1632-1642. 10, 2014.
63. Teplova M, Hafner M, Teplov D, Essig K, Tuschl T, Patel DJ. Structure-function studies of STAR family Quaking proteins bound to their in vivo RNA target sites. Genes Dev. 928-940. 27, 2013.
64. Maguire ML, Guler-Gane G, Nietlispach D, Raine ARC, Zorn AM, Standart N, Broadhurst RW. Solution structure and backbone dynamics of the Kh-Qua2 region of the Xenopus Star/Gsg Quaking protein. J Mol Biol. 265. 348, 2005.
65. Jacewicz A, Chico L, Smith P, Schwer B, Shuman S. Structural basis for recognition of intron branchpoint RNA by yeast Msl5 and selective effects of interfacial mutations on splicing of yeast pre-mRNAs. RNA. 401-414. 21, 2015.

WEBSITES (accessed in 2018)

- https://www.proteinatlas.org/ - April 7th
- https://www.uniprot.org/uniprot/ - April 11th
- http://bioinf.cs.ucl.ac.uk/psipred/ - April 10th
- http://www.compbio.dundee.ac.uk/jpred/ - April 13th
- https://www.proteinmodelportal.org/ - April 9th
- http://www.ebi.ac.uk/pdbsum – April 13th
- https://www.rcsb.org/ - April/May
- https://www.ebi.ac.uk/interpro/ - 3rd May
- https://blast.ncbi.nlm.nih.gov/Blast.cgi
- http://prdos.hgc.jp/cgi-bin/top.cgi
- http://www.cbs.dtu.dk/services/SignalP/
- https://toolkit.tuebingen.mpg.de/#/tools/hhpred – 28th April
- http://www.scfbio-iitd.res.in/bhageerathH+/ - 28th April
- http://raptorx.uchicago.edu/ – 28th April
- www.sbg.bio.ic.ac.uk/~phyre/ – 28th April
- http://www.reading.ac.uk/bioinf/IntFOLD/ – 28th April
- https://swissmodel.expasy.org/ - 28th April
- https://zhanglab.ccmb.med.umich.edu/I-TASSER/
- https://zhanglab.ccmb.med.umich.edu/QUARK/
- http://webs.iiitd.edu.in/raghava/pprint/index.html - 17th May
- http://ailab1.ist.psu.edu/RNABindRPlus/ - 18th May
- http://s.tartaglialab.com/page/catrapid_group - 18th May
- http://cib.cf.ocha.ac.jp/KYG/ - 20th May

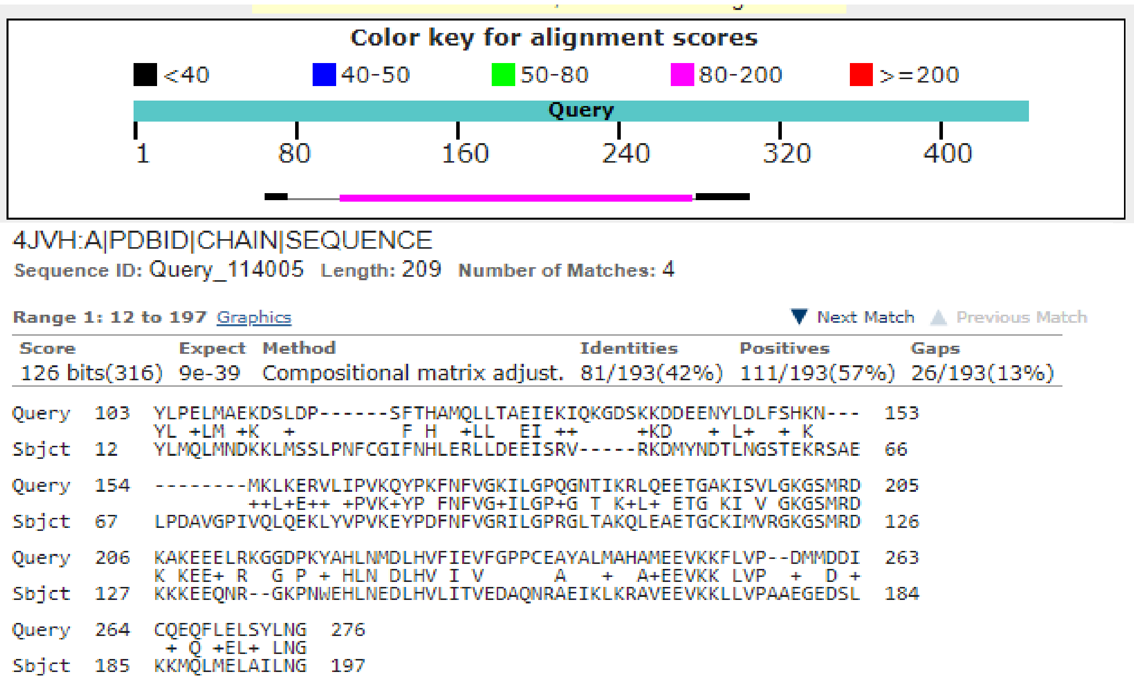
8. Appendices

8.1. PDB structures used as templates for homology modeling

This appendix contains four PDB structures which were used as templates for homology modeling by different servers. For each PDB, the output of the alignment between Sam68 and the PDB sequence is shown. As expected, all the sequences align in the KH domain region, ranging from 31% - 52% identity.

8.1.1. 4JVH

This is the structure of the star domain of quaking protein in complex with RNA in *Homo sapiens* [63]. This sequence includes the Qua1 and KH domain of Sam68. The alignment between Sam68 sequence and the PBD sequence is shown in Figure S1. The quaking protein is also a RNA-binding protein which plays a key role in myelinization, regulating mRNA stability, splicing, mRNA export and protein translation. This protein binds to this motif: 5'-NACUAAY-N$_{(1, 20)}$-UAAY-3'. The secondary structure of 4JVH is shown in Figure S2A. The RNA contact sites were consulted in PDBsum (Figure S2B) and a view of the three-dimensional structure is shown in Figure S2C.



Color key for alignment scores

⬛ <40 🟦 40-50 🟩 50-80 🟪 80-200 🟥 >=200

4JVH:A|PDBID|CHAIN|SEQUENCE

Sequence ID: Query_114005 Length: 209 Number of Matches: 4

Range 1: 12 to 197 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
126 bits(316)	9e-39	Compositional matrix adjust.	81/193(42%)	111/193(57%)	26/193(13%)

```
Query  103  YLPELMAEKDSLDP------SFTHAMQLLTAEIEKIQKGDSKKDDEENYLDLFSHKN---  153
            YL +LM +K  +         F H  +LL  EI ++     +KD   + L+  + K
Sbjct  12   YLMQLMNDKKLMSSLPNFCGIFNHLERLLDEEISRV-----RKDMYNDTLNGSTEKRSAE  66

Query  154  --------MKLKERVLIPVKQYPKFNFVGKILGPQGNTIKRLQEETGAKISVLGKGSMRD  205
                    ++L+E++ +PVK+YP FNFVG+ILGP+G T K+L+ ETG KI V GKGSMRD
Sbjct  67   LPDAVGPIVQLQEKLYVPVKEYPDFNFVGRILGPRGLTAKQLEAETGCKIMVRGKGSMRD  126

Query  206  KAKEEELRKGGDPKYAHLNMDLHVFIEVFGPPCEAYALMAHAMEEVKKFLVP--DMMDDI  263
            K KEE+ R  G P + HLN DLHV I V      A   +  A+EEVKK LVP  +  D +
Sbjct  127  KKKEEQNR--GKPNWEHLNEDLHVLITVEDAQNRAEIKLKRAVEEVKKLLVPAAEGEDSL  184

Query  264  CQEQFLELSYLNG  276
             + Q +EL+ LNG
Sbjct  185  KKMQLMELAILNG  197
```

Figure S1. Sequence alignment of Sam68 ("Query") and 4JVH ("Sbjct") using BLASTp.

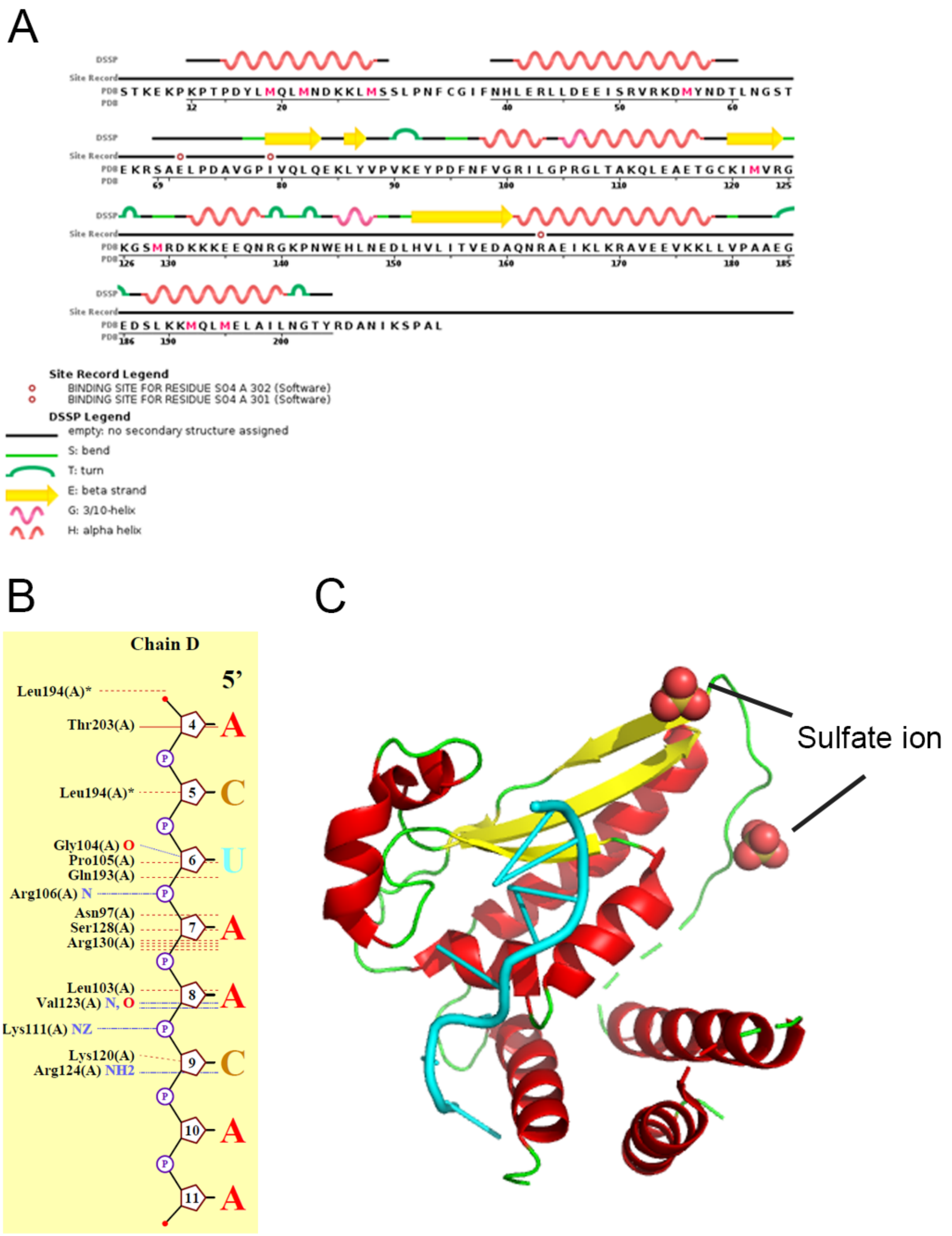
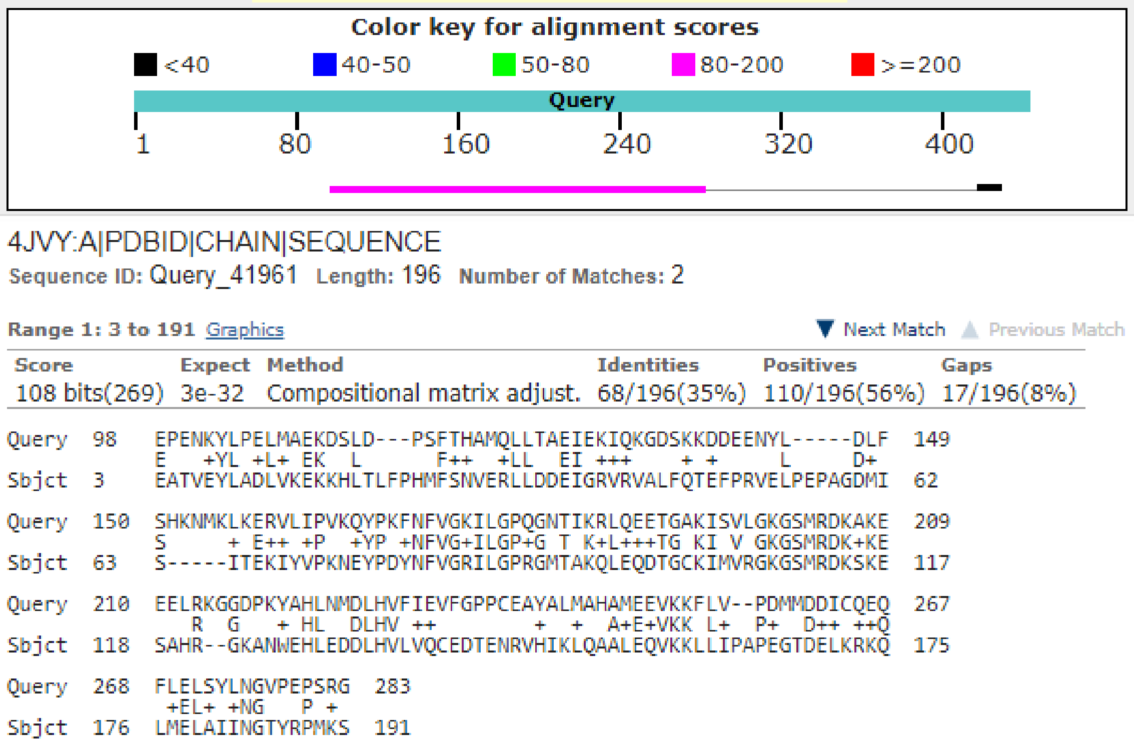


Figure S2. (A) Secondary structure of PDB 4JVH predicted by DSSP. (B) Protein-RNA contact sites. The pentagon shows backbone sugar and base-number, the P shows the phosphate group; the star indicates residue/water on plot more than once, the blue dashes show hydrogen bonds, and the red dashes show non-bonded contact to DNA/RNA (< 3.35Å). (C) Schematic view of the star domain of quaking protein, with the RNA motif showed in cyan.

8.1.2. 4JVY

Structure of the STAR (signal transduction and activation of RNA) domain of female germline-specific tumor suppressor GLD-1 bound to RNA in *Caenorhabditis elegans* [63]. This protein binds to 5'-UACUCAU-3' RNA sequence. Figure S3 shows the sequence alignment between Sam68 and the PDB sequences, whereas Figure S4 shows the secondary structure (S4A), the RNA contact sites (S4B) and the 3D structure (S4C).



Color key for alignment scores

⬛ <40 🟦 40-50 🟩 50-80 🟪 80-200 🟥 >=200

4JVY:A|PDBID|CHAIN|SEQUENCE

Sequence ID: Query_41961 Length: 196 Number of Matches: 2

Range 1: 3 to 191 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
108 bits(269)	3e-32	Compositional matrix adjust.	68/196(35%)	110/196(56%)	17/196(8%)

```
Query  98    EPENKYLPELMAEKDSLD---PSFTHAMQLLTAEIEKIQKGDSKKDDEENYL-----DLF  149
             E   +YL +L+ EK  L      F++  +LL  EI +++    + +     L     D+
Sbjct  3     EATVEYLADLVKEKKHLTLFPHMFSNVERLLDDEIGRVRVALFQTEFPRVELPEPAGDMI  62

Query  150   SHKNMKLKERVLIPVKQYPKFNFVGKILGPQGNTIKRLQEETGAKISVLGKGSMRDKAKE  209
             S     + E++ +P  +YP +NFVG+ILGP+G T K+L+++TG KI V GKGSMRDK+KE
Sbjct  63    S-----ITEKIYVPKNEYPDYNFVGRILGPRGMTAKQLEQDTGCKIMVRGKGSMRDKSKE  117

Query  210   EELRKGGDPKYAHLNMDLHVFIEVFGPPCEAYALMAHAMEEVKKFLV--PDMMDDICQEQ  267
                R  G   + HL  DLHV ++        +  +  A+E+VKK L+  P+  D++ ++Q
Sbjct  118   SAHR--GKANWEHLEDDLHVLVQCEDTENRVHIKLQAALEQVKKLLIPAPEGTDELKRKQ  175

Query  268   FLELSYLNGVPEPSRG  283
              +EL+ +NG   P +
Sbjct  176   LMELAIINGTYRPMKS  191
```

Figure S3. Sequence alignment of Sam68 (“Query”) and 4JVY (“Sbjct”) using BLASTp.

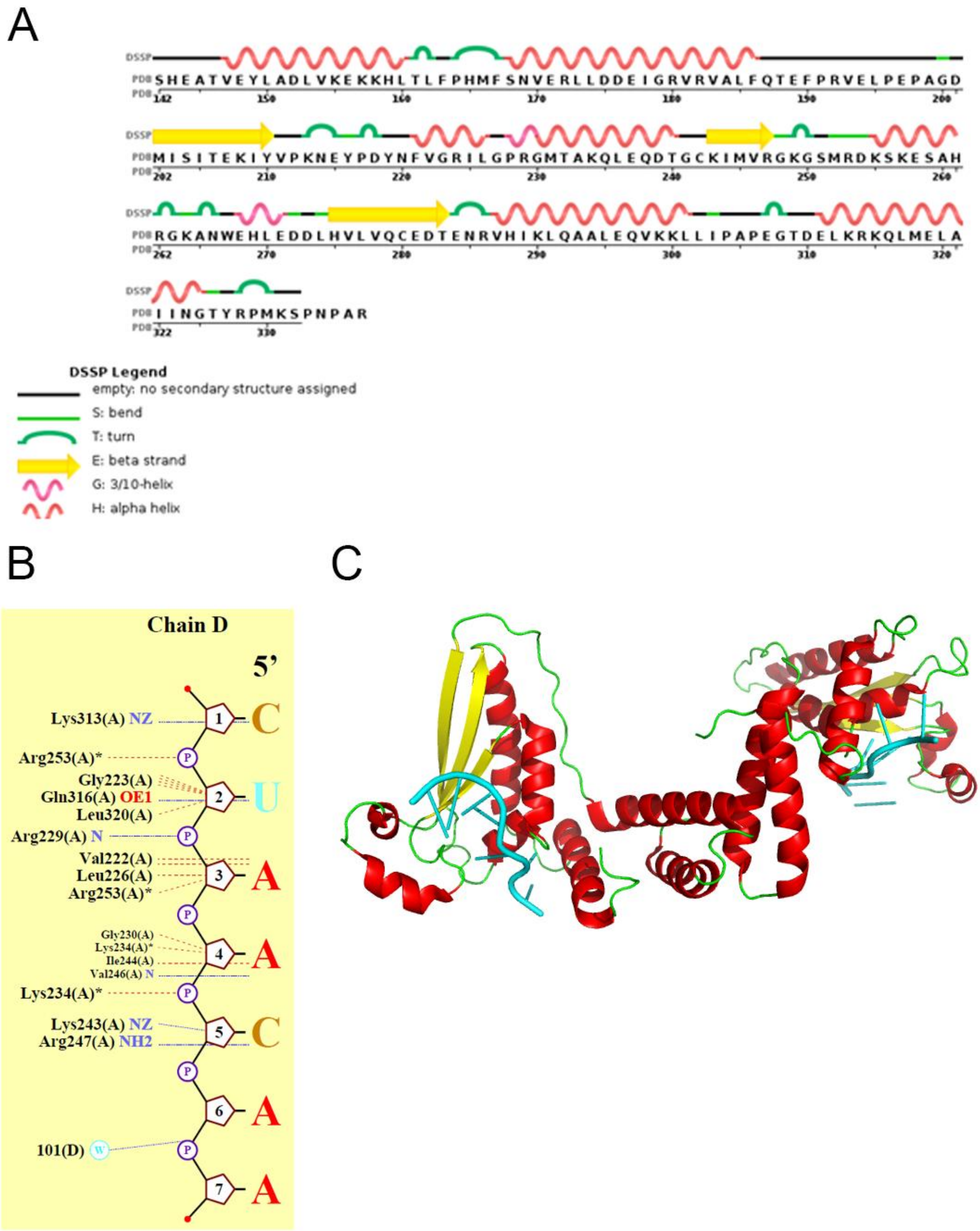
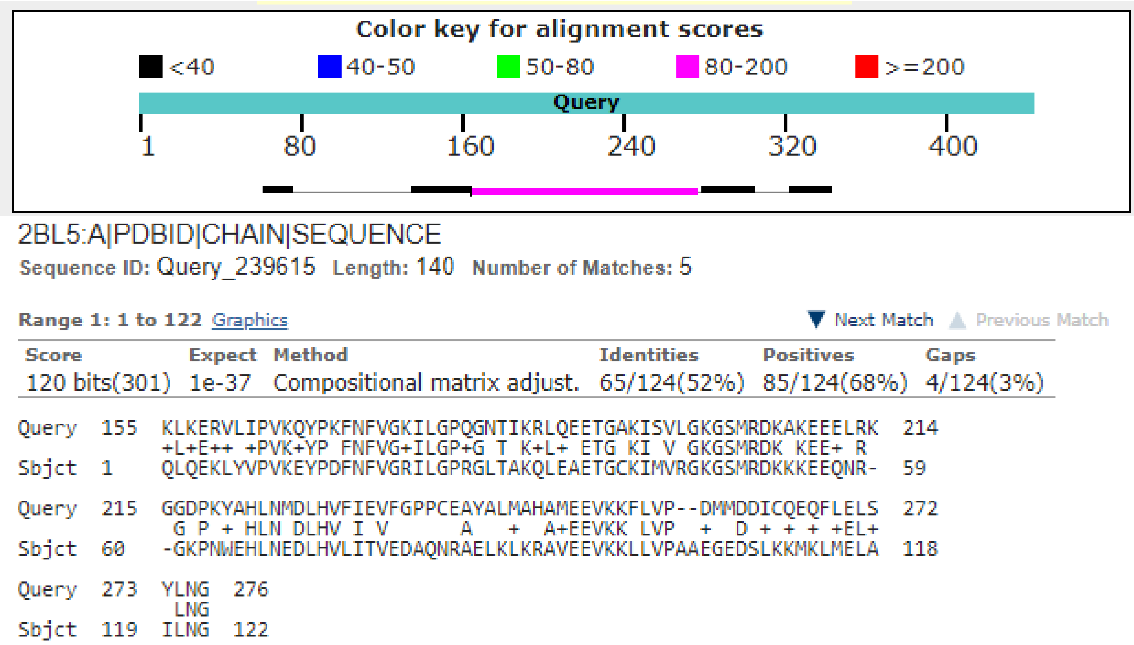

A
DSSP
PDB SHEATVEYLADLVKEKKHLTLFPHMFSNVERLLDDEIGRVRVALFQTEFPRVELPEPAGD
PDB 142 150 160 170 180 190 200
PDB MISITEKIYVPKNEYPDYNFVGRILGPRGMTAKQLEQDTGCKIMVRGKGSMRDKSKESAH
PDB 202 210 220 230 240 250 260
PDB RGKANWEHLEDDLHVLVQCEDTENRVHIKLQAALEQVKKLLIPAPEGTDELKRKQLMELA
PDB 262 270 280 290 300 310 320
PDB IINGTYRPMKSPNPAR
PDB 322 330
DSSP Legend
empty: no secondary structure assigned
S: bend
T: turn
E: beta strand
G: 3/10-helix
H: alpha helix
B
Chain D
5'
Lys313(A) NZ
Arg253(A)*
Gly223(A)
Gln316(A) OE1
Leu320(A)
Arg229(A) N
Val222(A)
Leu226(A)
Arg253(A)*
Gly230(A)
Lys234(A)*
Ile244(A)
Val246(A) N
Lys234(A)*
Lys243(A) NZ
Arg247(A) NH2
101(D)
W
P
1 C
2 U
3 A
4 A
5 C
6 A
7 A
C

Figure S4. (A) Secondary structure of PDB 4JVY predicted by DSSP. (B) Protein-RNA contact sites. The pentagon shows backbone sugar and base-number, the P shows the phosphate group; the star indicates residue/water on plot more than once, the blue dashes show hydrogen bonds, and the red dashes show non-bonded contact to DNA/RNA (< 3.35Å). (C) Schematic view of the STAR domain of GLD-1, with the RNA motif showed in cyan.

8.1.3. 2BL5

Solution structure of the KH-QUA2 region of the *Xenopus* STAR-GSG Quaking protein [64]. This protein binds to the same RNA sequence that quaking protein in *H. sapiens*. The sequence alignment is shown in Figure S5 and the secondary and 3D structures are shown in Figure S6.



Color key for alignment scores

■ <40   🟦 40-50   🟩 50-80   ■ 80-200   🟥 >=200

2BL5:A|PDBID|CHAIN|SEQUENCE

Sequence ID: Query_239615 Length: 140 Number of Matches: 5

Range 1: 1 to 122 Graphics

Score	Expect	Method	Identities	Positives	Gaps
120 bits(301)	1e-37	Compositional matrix adjust.	65/124(52%)	85/124(68%)	4/124(3%)

```
Query  155  KLKERVLIPVKQYPKFNFVGKILGPQGNTIKRLQEETGAKISVLGKGSMRDKAKEEELRK  214
            +L+E++ +PVK+YP FNFVG+ILGP+G T K+L+ ETG KI V GKGSMRDK KEE+ R
Sbjct  1    QLQEKLYVPVKEYPDFNFVGRILGPRGLTAKQLEAETGCKIMVRGKGSMRDKKKEEQNR-  59

Query  215  GGDPKYAHLNMDLHVFIEVFGPPCEAYALMAHAMEEVKKFLVP--DMMDDICQEQFLELS  272
             G P + HLN DLHV I V      A   +  A+EEVKK LVP  +  D + + + +EL+
Sbjct  60   -GKPNWEHLNEDLHVLITVEDAQNRAELKLKRAVEEVKKLLVPAAEGEDSLKKMKLMELA  118

Query  273  YLNG  276
             LNG
Sbjct  119  ILNG  122
```

Figure S5. Sequence alignment of Sam68 (“Query”) and 2BL5 (“Sbjct”) using BLASTp.

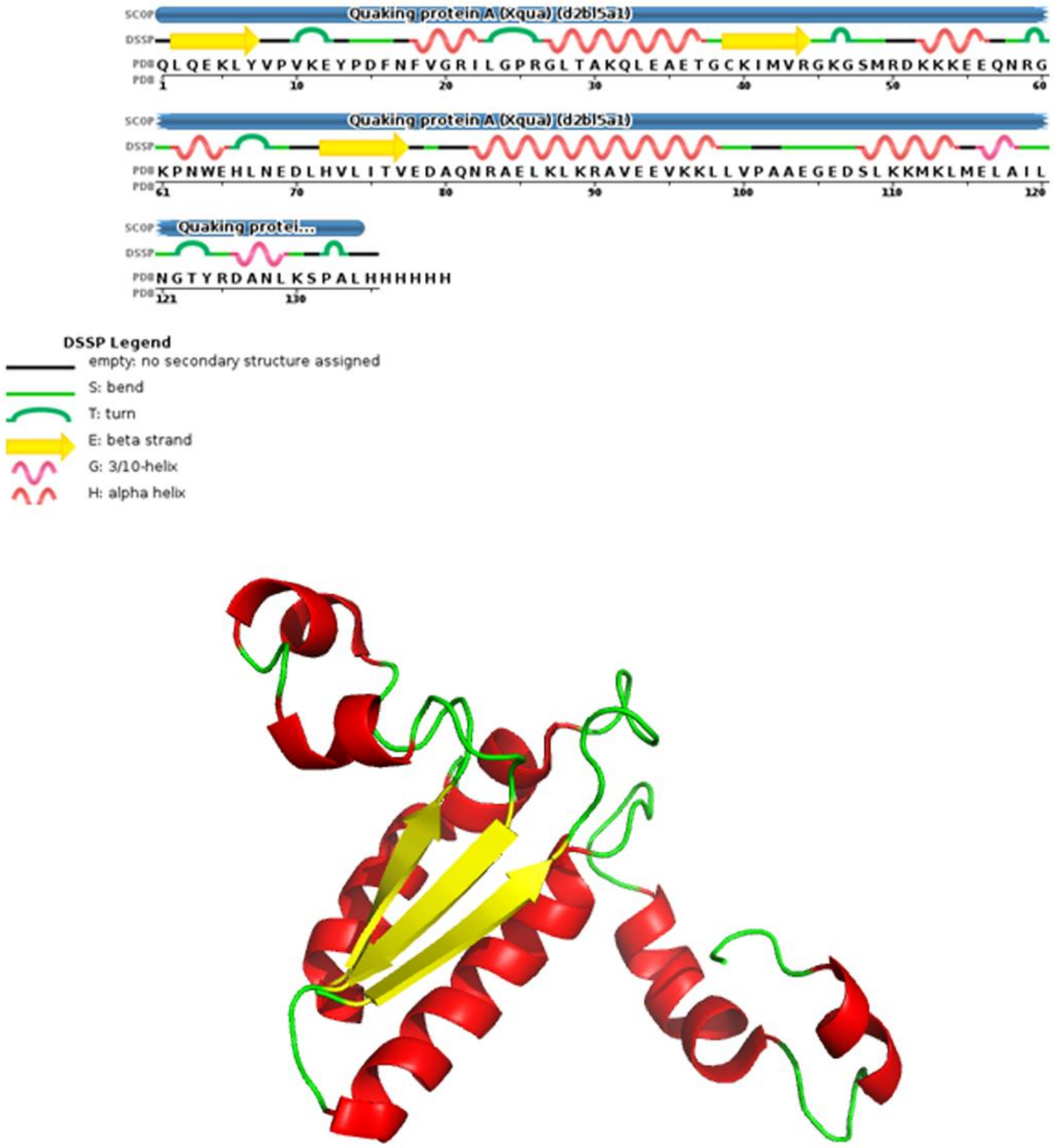
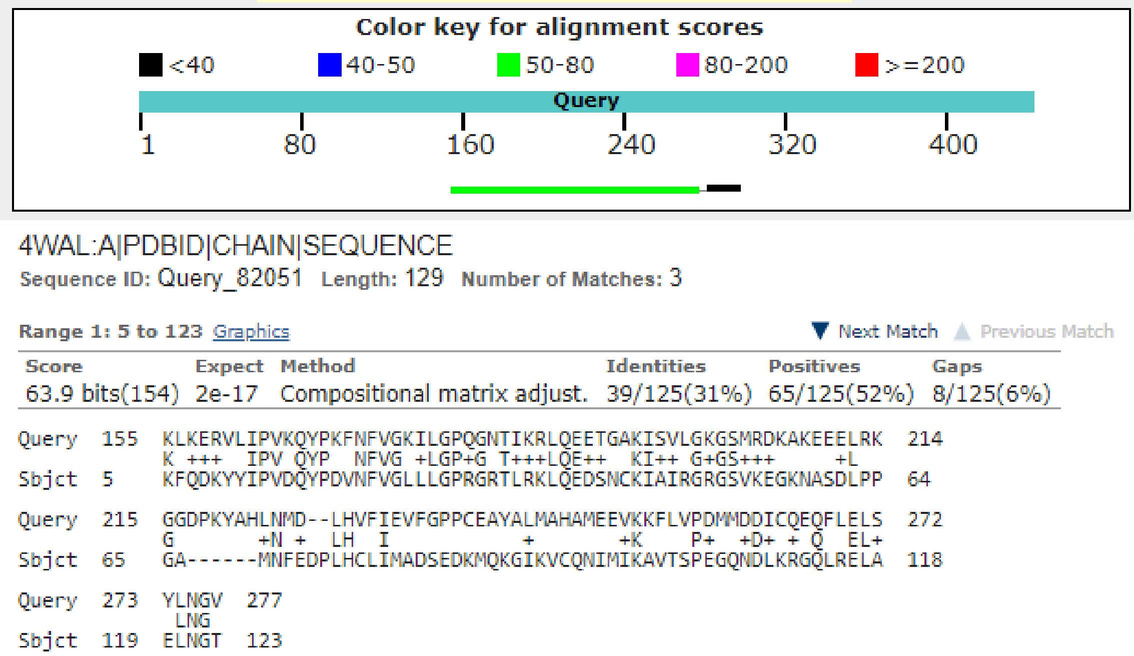

SCOP
Quaking protein A (Xqua) (d2bl5a1)
DSSP
PDB QLQEKLYVPVKEYPDFNFVGRILGPRGLTAKQLEAETGCKIMVRGKGSMRDKKKEEQNRG
PDB 1 10 20 30 40 50 60
SCOP
Quaking protein A (Xqua) (d2bl5a1)
DSSP
PDB KPNWEHLNEDLHVLITVEDAQNRAELKLKRAVEEVKKLLVPAAEGEDSLKKMKLMELAIL
PDB 61 70 80 90 100 110 120
SCOP
Quaking protei...
DSSP
PDB NGTYRDANLKSPALHHHHHH
PDB 121 130
DSSP Legend
empty: no secondary structure assigned
S: bend
T: turn
E: beta strand
G: 3/10-helix
H: alpha helix

Figure S6. Secondary structure of PDB 2BL5 and the schematic view of the 3D structure.

8.1.4. 4WAL

Crystal structure of selenomethionine Msl5 protein in complex with RNA at 2.2 Å [65]. This is an RNA binding protein also involved in mRNA splicing in *Saccharomyces cerevisiae*. The alignment of this protein with Sam68 is shown in Figure S7; this reported the lowest percentage of identity (31%). Figure S8A shows the secondary structure of the PDB, Figure S8B illustrates the RNA contact sites consulted in PDBsum, and Figure S8C shows the 3D representation of the structure.



4WAL:A|PDBID|CHAIN|SEQUENCE

Sequence ID: Query_82051 Length: 129 Number of Matches: 3

Range 1: 5 to 123 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
63.9 bits(154)	2e-17	Compositional matrix adjust.	39/125(31%)	65/125(52%)	8/125(6%)

```
Query  155  KLKERVLIPVKQYPKFNFVGKILGPQGNTIKRLQEETGAKISVLGKGSMRDKAKEEELRK  214
            K +++  IPV QYP  NFVG +LGP+G T+++LQE++  KI++ G+GS+++     +L
Sbjct  5    KFQDKYYIPVDQYPDVNFVGLLLGPRGRTLRKLQEDSNCKIAIRGRGSVKEGKNASDLPP  64

Query  215  GGDPKYAHLNMD--LHVFIEVFGPPCEAYALMAHAMEEVKKFLVPDMMDDICQEQFLELS  272
            G       +N +  LH  I           +       +K    P+  +D+ + Q  EL+
Sbjct  65   GA------MNFEDPLHCLIMADSEDKMQKGIKVCQNIMIKAVTSPEGQNDLKRGQLRELA  118

Query  273  YLNGV  277
             LNG
Sbjct  119  ELNGT  123
```

Figure S7. Sequence alignment of Sam68 (“Query”) and 4WAL (“Sbjct”) using BLASTp.

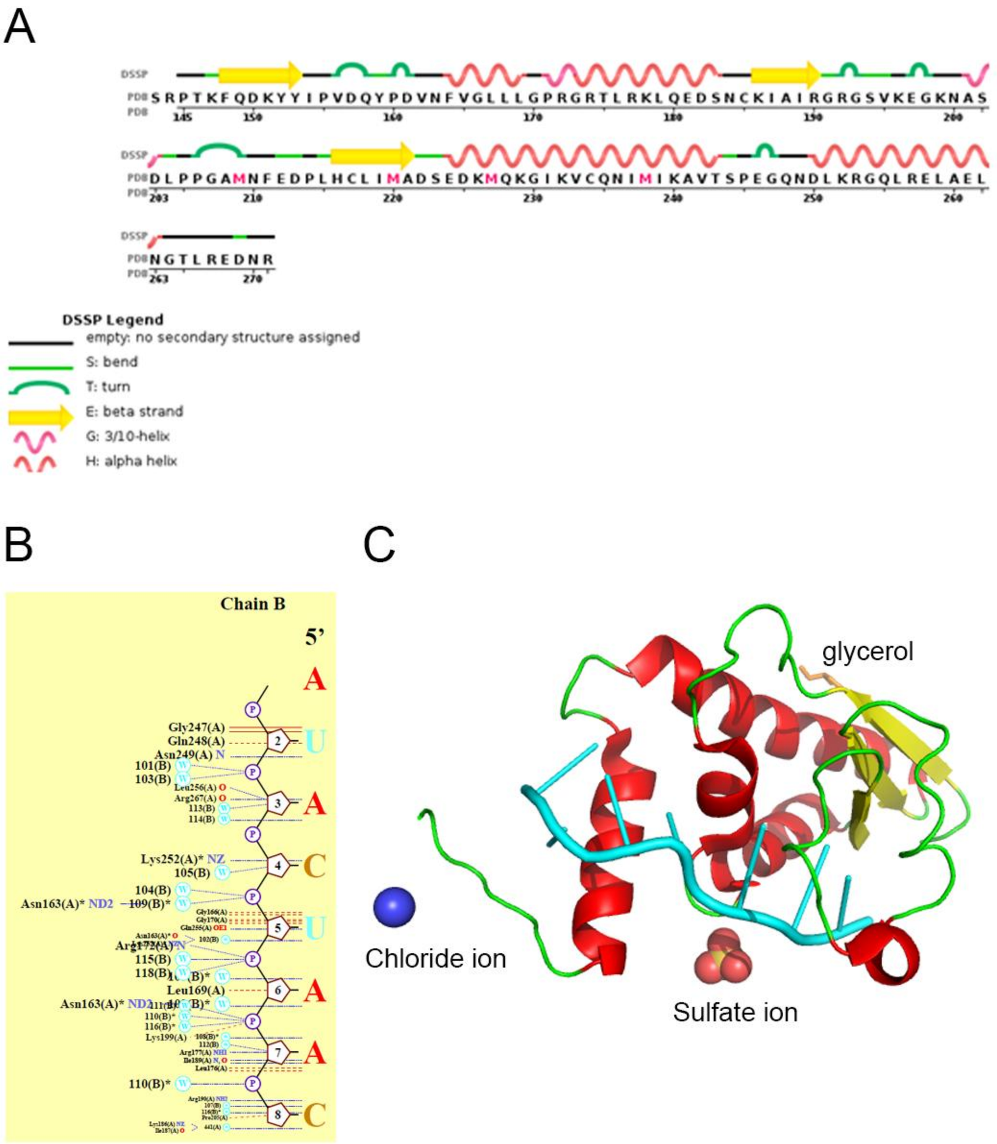


Figure S8. (A) Secondary structure of PDB 4WAL predicted by DSSP. (B) Protein-RNA contact sites. The pentagon shows backbone sugar and base-number, the P shows the phosphate group; the star indicates residue/water on plot more than once, the blue dashes show hydrogen bonds, and the red dashes show non-bonded contact to DNA/RNA (< 3.35Å). (C) Schematic view of Msl5 protein with the RNA motif showed in cyan.

8.2. SMN2 transcripts

- This variant (d) represents the longest transcript and encodes the longest isoform (d).

>NM_017411.3 Homo sapiens survival of motor neuron 2, centromeric (SMN2), transcript variant d, mRNA
CCACAAAUGUGGGAGGGCGAUAACCACUCGUAGAAAGCGUGAGAAGUUACUACAAGCGGUCCUCCCGGCC
ACCGUACUGUUCCGCUCCCAGAAGCCCCGGGCGGCGGAAGUCGUCACUCUUAAGAAGGG**A**CGGGGCCCCA
CGCUGCGCACCCGCGGGUUUGCUAUGGCGAUGAGCAGCGGCGGCAGUGGUGGCG**G**CGUCCCGGAGCAGGA
GGAUUCCGUGCUGUUCCGGCGCGGCACAGGCCAGAGCGAUGAUUCUGACAUUUGGGAUGAUACAGCACUG
AUAAAAGCAUAUGAUAAAGCUGUGGCUUCAUUUAAGCAUGCUCUAAAGAAUGGUGACAUUUGUGAAACUU
CGGGUAAACCAAAAACCACACCUAAAAGAAAACCUGCUAAGAAGAAUAAAAGCCAAAAGAAGAAUACUGC
AGCUUCCUUACAACAGUGGAAAGUUGGGGACAAAUGUUCUGCCAUUUGGUCAGAAGACGGUUGCAUUUAC
CCAGCUACCAUUGCUUCAAUUGAUUUUAAGAGAGAAACCUGUGUUGUGGUUUACACUGGAUAUGGAAAUA
GAGAGGAGCAAAAUCUGUCCGAUCUACUUUCCCCAAUCUGUGAAGUAGCUAAUAAUAUAGAACAAAAUGC
UCAAGAGAAUGAAAAUGAAAGCCAAGUUUCAACAGAUGAAAGUGAGAACUCCAGGUCUCCUGGAAAUAAA
UCAGAUAACAUCAAGCCCAAAUCUGCUCCAUGGAACUCUUUUCUCCCUCCACCACCCCCCAUGCCAGGGC
CAAGACUGGGACCAGGAAAGCCAGGUCUAAAAUUCAAUGGCCCACCACCGCCACCGCCACCACCACCACC
CCACUUACUAUCAUGCUGGCUGCCUCCAUUUCCUUCUGGACCACCAAUAAUUCCCCCACCACCUCCCAUA
UGUCCAGAUUCUCUUGAUGAUGCUGAUGCUUUGGGAAGUAUGUUAAUUUCAUGGUACAUGAGUGGCUAUC
AUACUGGCUAUUAUAUGGGUUUUAGACAAAAUCAAAAAGAAGGAAGGUGCUCACAUUCCUUAAAUUAAGG
AGAAAUGCUGGCAUAGAGCAGCACUAAAUGACACCACUAAAGAAACGAUCAGACAGAUCUGGAAUGUGAA
GCGUUAUAGAAGAUAACUGGCCUCAUUUCUUCAAAAUAUCAAGUGUUGGGAAAGAAAAAAGGAAGUGGAA
UGGGUAACUCUUCUUGAUUAAAAGUUAUGUAAUAACCAAAUGCAAUGUGAAAUAUUUUACUGGACUCUAU
UUUGAAAAACCAUCUGUAAAAGACUGAGGUGGGGGUGGGAGGCCAGCACGGUGGUGAGGCAGUUGAGAAA
AUUUGAAUGUGGAUUAGAUUUUGAAUGAUAUUGGAUAAUUAUUGGUAAUUUUAUGAGCUGUGAGAAGGGU
GUUGUAGUUUAUAAAAGACUGUCUUAAUUUGCAUACUUAAGCAUUUAGGAAUGAAGUGUUAGAGUGUCUU
AAAAUGUUUCAAAUGGUUUAACAAAAUGUAUGUGAGGCGUAUGUGGCAAAAUGUUACAGAAUCUAACUGG
UGGACAUGGCUGUUCAUUGUACUGUUUUUUUCUAUCUUCUAUAUGUUUAAAAGUAUAUAAUAAAAAUAUU
UAAUUUUUUUUUAAAUUAAAAAAA

- This variant (a) lacks an alternate exon in the 3' CDS compared to variant d. The resulting protein (isoform a) is shorter and has a distinct C-terminus compared to isoform d. This variant is thought to be the predominant transcript produced by this copy of the gene.

>NM_022875.2 Homo sapiens survival of motor neuron 2, centromeric (SMN2), transcript variant a, mRNA
CCACAAAUGUGGGAGGGCGAUAACCACUCGUAGAAAGCGUGAGAAGUUACUACAAGCGGUCCUCCCGGCC
ACCGUACUGUUCCGCUCCCAGAAGCCCCGGGCGGCGGAAGUCGUCACUCUUAAGAAGGGACGGGGCCCCA
CGCUGCGCACCCGCGGGUUUGCUAUGGCGAUGAGCAGCGGCGGCAGUGGUGGCGGCGUCCCGGAGCAGGA
GGAUUCCGUGCUGUUCCGGCGCGGCACAGGCCAGAGCGAUGAUUCUGACAUUUGGGAUGAUACAGCACUG
AUAAAAGCAUAUGAUAAAGCUGUGGCUUCAUUUAAGCAUGCUCUAAAGAAUGGUGACAUUUGUGAAACUU
CGGGUAAACCAAAAACCACACCUAAAAGAAAACCUGCUAAGAAGAAUAAAAGCCAAAAGAAGAAUACUGC
AGCUUCCUUACAACAGUGGAAAGUUGGGGACAAAUGUUCUGCCAUUUGGUCAGAAGACGGUUGCAUUUAC
CCAGCUACCAUUGCUUCAAUUGAUUUUAAGAGAGAAACCUGUGUUGUGGUUUACACUGGAUAUGGAAAUA
GAGAGGAGCAAAAUCUGUCCGAUCUACUUUCCCCAAUCUGUGAAGUAGCUAAUAAUAUAGAACAAAAUGC
UCAAGAGAAUGAAAAUGAAAGCCAAGUUUCAACAGAUGAAAGUGAGAACUCCAGGUCUCCUGGAAAUAAA
UCAGAUAACAUCAAGCCCAAAUCUGCUCCAUGGAACUCUUUUCUCCCUCCACCACCCCCCAUGCCAGGGC
CAAGACUGGGACCAGGAAAGCCAGGUCUAAAAUUCAAUGGCCCACCACCGCCACCGCCACCACCACCACC
CCACUUACUAUCAUGCUGGCUGCCUCCAUUUCCUUCUGGACCACCAAUAAUUCCCCCACCACCUCCCAUA
UGUCCAGAUUCUCUUGAUGAUGCUGAUGCUUUGGGAAGUAUGUUAAUUUCAUGGUACAUGAGUGGCUAUC
AUACUGGCUAUUAUAUGGAAAUGCUGGCAUAGAGCAGCACUAAAUGACACCACUAAAGAAACGAUCAGAC

```
AGAUCUGGAAUGUGAAGCGUUAUAGAAGAUAACUGGCCUCAUUUCUUCAAAAUAUCAAGUGUUGGGAAAG
AAAAAAGGAAGUGGAAUGGGUAACUCUUCUUGAUUAAAAGUUAUGUAAUAACCAAAUGCAAUGUGAAAUA
UUUUACUGGACUCUAUUUUGAAAAACCAUCUGUAAAAGACUGAGGUGGGGGUGGGAGGCCAGCACGGUGG
UGAGGCAGUUGAGAAAAUUUGAAUGUGGAUUAGAUUUUGAAUGAUAUUGGAUAAUUAUUGGUAAUUUUAU
GAGCUGUGAGAAGGGUGUUGUAGUUUAUAAAAGACUGUCUUAAUUUGCAUACUUAAGCAUUUAGGAAUGA
AGUGUUAGAGUGUCUUAAAAUGUUUCAAAUGGUUUAACAAAAUGUAUGUGAGGCGUAUGUGGCAAAAUGU
UACAGAAUCUAACUGGUGGACAUGGCUGUUCAUUGUACUGUUUUUUUCUAUCUUCUAUAUGUUUAAAAGU
AUAUAAUAAAAAUAUUUAAUUUUUUUUUAAAUUAAAAAAA
```

- This variant (c) lacks two alternate exons in the 3' CDS compared to variant d. The resulting protein (isoform c) is shorter and has a distinct C-terminus compared to isoform d.

>NM_022877.2 Homo sapiens survival of motor neuron 2, centromeric (SMN2), transcript variant c, mRNA

```
CCACAAAUGUGGGAGGGCGAUAACCACUCGUAGAAAGCGUGAGAAGUUACUACAAGCGGUCCUCCCGGCC
ACCGUACUGUUCCGCUCCCAGAAGCCCCGGGCGGCGGAAGUCGUCACUCUUAAGAAGGGACGGGGCCCCA
CGCUGCGCACCCGCGGGUUUGCUAUGGCGAUGAGCAGCGGCGGCAGUGGUGGCGGCGUCCCGGAGCAGGA
GGAUUCCGUGCUGUUCCGGCGCGGCACAGGCCAGAGCGAUGAUUCUGACAUUUGGGAUGAUACAGCACUG
AUAAAAGCAUAUGAUAAAGCUGUGGCUUCAUUUAAGCAUGCUCUAAAGAAUGGUGACAUUUGUGAAACUU
CGGGUAAACCAAAAACCACACCUAAAAGAAAACCUGCUAAGAAGAAUAAAAGCCAAAAGAAGAAUACUGC
AGCUUCCUUACAACAGUGGAAAGUUGGGGACAAAUGUUCUGCCAUUUGGUCAGAAGACGGUUGCAUUUAC
CCAGCUACCAUUGCUUCAAUUGAUUUUAAGAGAGAAACCUGUGUUGUGGUUUACACUGGAUAUGGAAAUA
GAGAGGAGCAAAAUCUGUCCGAUCUACUUUCCCCAAUCUGUGAAGUAGCUAAUAAUAUAGAACAAAAUGC
UCAAGAGAAUGAAAAUGAAAGCCAAGUUUCAACAGAUGAAAGUGAGAACUCCAGGUCUCCUGGAAAUAAA
UCAGAUAACAUCAAGCCCAAAUCUGCUCCAUGGAACUCUUUUCUCCCUCCACCACCCCCCAUGCCAGGGC
CAAGACUGGGACCAGGAAAGAUAAUUCCCCCACCACCUCCCAUAUGUCCAGAUUCUCUUGAUGAUGCUGA
UGCUUUGGGAAGUAUGUUAAUUUCAUGGUACAUGAGUGGCUAUCAUACUGGCUAUUAUAUGGAAAUGCUG
GCAUAGAGCAGCACUAAAUGACACCACUAAAGAAACGAUCAGACAGAUCUGGAAUGUGAAGCGUUAUAGA
AGAUAACUGGCCUCAUUUCUUCAAAAUAUCAAGUGUUGGGAAAGAAAAAAGGAAGUGGAAUGGGUAACUC
UUCUUGAUUAAAAGUUAUGUAAUAACCAAAUGCAAUGUGAAAUAUUUUACUGGACUCUAUUUUGAAAAAC
CAUCUGUAAAAGACUGAGGUGGGGGUGGGAGGCCAGCACGGUGGUGAGGCAGUUGAGAAAAUUUGAAUGU
GGAUUAGAUUUUGAAUGAUAUUGGAUAAUUAUUGGUAAUUUUAUGAGCUGUGAGAAGGGUGUUGUAGUUU
AUAAAAGACUGUCUUAAUUUGCAUACUUAAGCAUUUAGGAAUGAAGUGUUAGAGUGUCUUAAAAUGUUUC
AAAUGGUUUAACAAAAUGUAUGUGAGGCGUAUGUGGCAAAAUGUUACAGAAUCUAACUGGUGGACAUGGC
UGUUCAUUGUACUGUUUUUUUCUAUCUUCUAUAUGUUUAAAAGUAUAUAAUAAAAAUAUUUAAUUUUUUU
UUAAAUUAAAAAAA
```

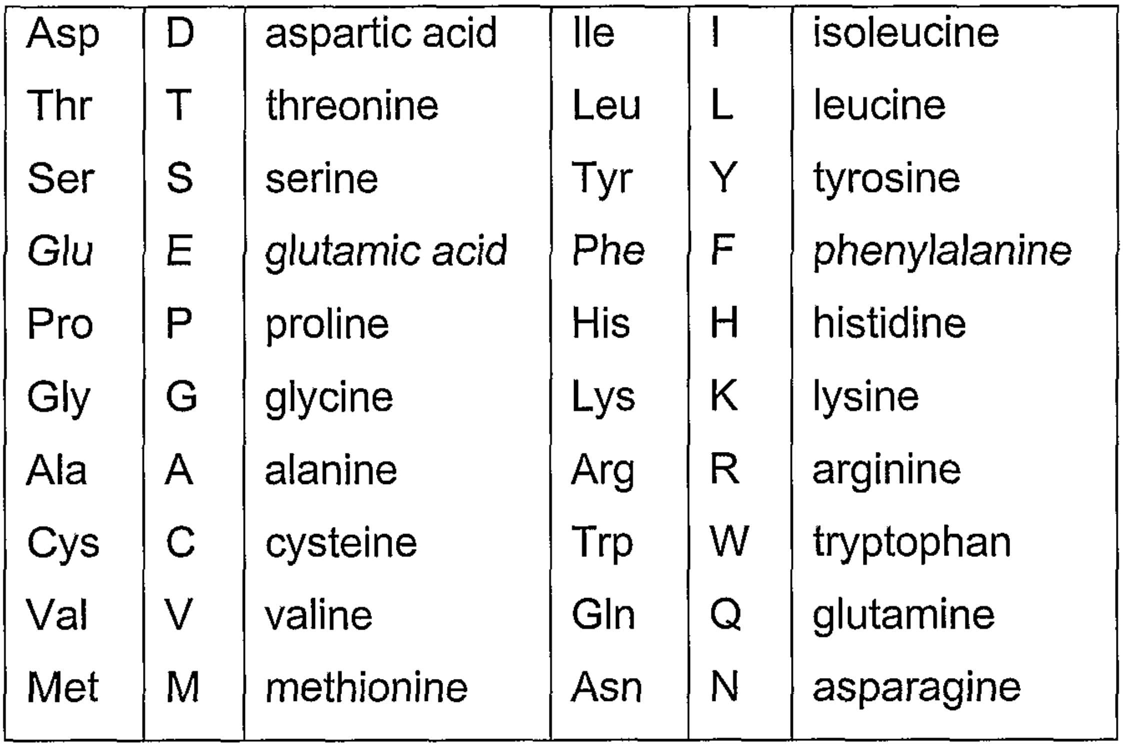
- This variant (b) lacks an alternate in-frame exon in the 3' CDS compared to variant d. The resulting protein (isoform b) is shorter but has the same N- and C- termini compared to isoform d.

>NM_022876.2 Homo sapiens survival of motor neuron 2, centromeric (SMN2), transcript variant b, mRNA

```
CCACAAAUGUGGGAGGGCGAUAACCACUCGUAGAAAGCGUGAGAAGUUACUACAAGCGGUCCUCCCGGCC
ACCGUACUGUUCCGCUCCCAGAAGCCCCGGGCGGCGGAAGUCGUCACUCUUAAGAAGGGACGGGGCCCCA
CGCUGCGCACCCGCGGGUUUGCUAUGGCGAUGAGCAGCGGCGGCAGUGGUGGCGGCGUCCCGGAGCAGGA
GGAUUCCGUGCUGUUCCGGCGCGGCACAGGCCAGAGCGAUGAUUCUGACAUUUGGGAUGAUACAGCACUG
AUAAAAGCAUAUGAUAAAGCUGUGGCUUCAUUUAAGCAUGCUCUAAAGAAUGGUGACAUUUGUGAAACUU
CGGGUAAACCAAAAACCACACCUAAAAGAAAACCUGCUAAGAAGAAUAAAAGCCAAAAGAAGAAUACUGC
AGCUUCCUUACAACAGUGGAAAGUUGGGGACAAAUGUUCUGCCAUUUGGUCAGAAGACGGUUGCAUUUAC
CCAGCUACCAUUGCUUCAAUUGAUUUUAAGAGAGAAACCUGUGUUGUGGUUUACACUGGAUAUGGAAAUA
GAGAGGAGCAAAAUCUGUCCGAUCUACUUUCCCCAAUCUGUGAAGUAGCUAAUAAUAUAGAACAAAAUGC
UCAAGAGAAUGAAAAUGAAAGCCAAGUUUCAACAGAUGAAAGUGAGAACUCCAGGUCUCCUGGAAAUAAA
```

```
UCAGAUAACAUCAAGCCCAAAUCUGCUCCAUGGAACUCUUUUCUCCCUCCACCACCCCCCAUGCCAGGGC
CAAGACUGGGACCAGGAAAGAUAAUUCCCCCACCACCUCCCAUAUGUCCAGAUUCUCUUGAUGAUGCUGA
UGCUUUGGGAAGUAUGUUAAUUUCAUGGUACAUGAGUGGCUAUCAUACUGGCUAUUAUAUGGGUUUUAGA
CAAAAUCAAAAAGAAGGAAGGUGCUCACAUUCCUUAAAUUAAGGAGAAAUGCUGGCAUAGAGCAGCACUA
AAUGACACCACUAAAGAAACGAUCAGACAGAUCUGGAAUGUGAAGCGUUAUAGAAGAUAACUGGCCUCAU
UUCUUCAAAAUAUCAAGUGUUGGGAAAGAAAAAAGGAAGUGGAAUGGGUAACUCUUCUUGAUUAAAAGUU
AUGUAAUAACCAAAUGCAAUGUGAAAUAUUUUACUGGACUCUAUUUUGAAAAACCAUCUGUAAAAGACUG
AGGUGGGGGUGGGAGGCCAGCACGGUGGUGAGGCAGUUGAGAAAAUUUGAAUGUGGAUUAGAUUUUGAAU
GAUAUUGGAUAAUUAUUGGUAAUUUUAUGAGCUGUGAGAAGGGUGUUGUAGUUUAUAAAAGACUGUCUUA
AUUUGCAUACUUAAGCAUUUAGGAAUGAAGUGUUAGAGUGUCUUAAAAUGUUUCAAAUGGUUUAACAAAA
UGUAUGUGAGGCGUAUGUGGCAAAAUGUUACAGAAUCUAACUGGUGGACAUGGCUGUUCAUUGUACUGUU
UUUUUCUAUCUUCUAUAUGUUUAAAAGUAUAUAAUAAAAAUAUUUAAUUUUUUUUUAAAUUAAAAAAA
```

8.3. Amino acid nomenclature



Asp	D	aspartic acid	Ile	I	isoleucine
Thr	T	threonine	Leu	L	leucine
Ser	S	serine	Tyr	Y	tyrosine
Glu	*E*	*glutamic acid*	*Phe*	*F*	*phenylalanine*
Pro	P	proline	His	H	histidine
Gly	G	glycine	Lys	K	lysine
Ala	A	alanine	Arg	R	arginine
Cys	C	cysteine	Trp	W	tryptophan
Val	V	valine	Gln	Q	glutamine
Met	M	methionine	Asn	N	asparagine