In silico investigations of mRNA binding specificity by human Quaking protein

Apoorv Alawada

A dissertation submitted for the partial fulfilment of
BS-MS dual degree in Science

Indian Institute of Science Education and Research Mohali

April 2019

Certificate of Examination

This to certify that the dissertation titled **"In silico investigations of mRNA binding specificity by human Quaking protein"** submitted by **Mr. Apoorv Alawada** (Reg. No. MS14054) for the partial fulfilment of BS-MS dual degree programme of the Institute, has been examined by the thesis committee duly appointed by the Institute. The committee finds the work done by the candidate satisfactory and recommends that the report be accepted.

Dr. Ujjal K Gautam Dr. Shashi Bhushan Pandit Dr. Monika Sharma

(Supervisor)

Dated: April 26, 2019

Declaration

The work presented in the dissertation has been carried by me under the guidance of Dr. Monika Sharma at the Indian Institute of Science Education and Research Mohali.

This work has not been submitted in part or in full for a degree, or diploma, or a fellowship to any other University or Institute. Whenever contribution of others are involved, every effort is made to indicate this clearly, with due acknowledgement of collaborative research and discussions. This thesis is a bonafide record of original work done by me and all sources listed within have been detailed in bibliography.

Apoorv Alawada

(Candidate)

Dated: April 26, 2019

In my capacity as the supervisor of the candidate's project work, I certify that the above statements by the candidate are true to the best of my knowledge

Dr. Monika Sharma

(Supervisor)

Acknowledgement

I would like to acknowledge my project guide Dr.Monika Sharma and my friend Jorawar Singh for their guidance and immense support during the project.

List of figures

Notations

$G_{complex}$ = Gibbs free energy for protein ligand complex

$\langle E_{MM} \rangle$ = average molecular mechanics potential energy in a vacuum.

$\phi(r)$ = electrostatic potential

$\varepsilon(r)$ = dielectric constant

$\rho^f(r)$ = fixed charge density

γ = coefficient related to surface tension of the solvent

b = fitting parameter

Abstract

Understanding the recognition mechanism of STAR family protein with RNA has been a challenging task in molecular and computational biology. Previous it has been reported that the binding shows a preference towards a specific codon sequence which is YUAAY. Here, to understand it better we try to mutate it and simulate it for 5ns and then study root mean square fluctuation, RMSD and binding energy for each mutation with QKI protein.

Contents

1: Introduction

1.1 mRNA complexation with protein

Mammalian Quaking (QKI) STAR (Signal Transduction and Activation of Ribonucleic acid) protein are evolutionarily conserved RNA-binding proteins, which post-transcriptionally regulate target genes essential for developmental processes and myelination recognition elements insights.[1] STAR protein family has a highly conserved domain of approximately 200 amino acids composed of QUA1, K homology (KH), and QUA2 motifs which makes them bound in high-affinity with common in vivo RNA targets containing YUAAY RNA recognition elements (RREs).[2-6] The QUA1 domain can mediate protein dimerization,[7,8] while the QUA2 domain is involved together with the KH domain in RNA binding.[9]

QKI have been reported to be related with several neurological disorders like schizophrenia, ataxia, multiple sclerosis and also cancer.[2,10] Other STAR family members like Splicing Factor 1 (SF1) mediate the intron recognition during spliceosome assembly, [11-13] QR(Quaking) proteins found to be critical in modulating development processes such as mammalian spermatogenesis and metazoan central nervous system development (GLD-1 in worms), sperm-to-oogenesis in hermaphrodites (GLD-1).[14-20] This all invokes the need for us to better understand the binding specificity.

Site1	ACACACUAACCU	906-918
Site2	GAAAAAUAACCA	923-935
Site3	GUCUAAUAAUGU	1060-1072
Site4	CCACAAUAACGU	1273-1285
Site5	AAAUAAUAAACU	1457-1469

Fig.1.1 Nucleotide sequence and position of Qk1 binding sites in the 3'-UTR of myelin basic protein mRNA [9]

Previously in an experimental study the relative affinity of each site in figure 1 with respect to QKI STAR found out 1 with highest affinity then 3 as second most affinity, then 2 and 4 bind equivalently well relatively. Computational study have been done previously for

wild type in this project we try to make 5 mutations with respect to wild type sequence and simulate it for 5ns and do comparative study.

1.2 Simulation

Computer simulations act as a bridge between microscopic length and timescales and the macroscopic world of the laboratory i.e. we provide a guess at the interactions between molecules, and obtain quite exact predictions of bulk properties. By simulations the hidden details behind the measurements can be explored. MD simulations solve Newton's equations of motion for a system of N interacting atoms:

$$m_i \frac{\partial^2 \mathbf{r}_i}{\partial t^2} = \mathbf{F}_i,\ i = 1 \ldots N. \tag{1.1}$$

The forces are the negative derivatives of a potential function $V(r_1, r_2, \ldots, r_N)$:

$$\mathbf{F}_i = -\frac{\partial V}{\partial \mathbf{r}_i} \tag{1.2}$$

The equations are solved simultaneously in small time steps. The system is followed for some time, taking care that the temperature and pressure remain at the required values, and the coordinates are written to an output file at regular intervals. The coordinates as a function of time represent a trajectory of the system. After initial changes, the system will usually reach an equilibrium state. By averaging over an equilibrium trajectory, many macroscopic properties can be extracted from the output file.

The potential energy function of a macro molecular system is a very complex hyperspace in a large number of dimensions. It has one deepest point as the global minimum and a very large number of local minima, where all derivatives of the potential energy function with respect to the coordinates are zero and all second derivatives are non-negative. In between the local minima there are saddle points, where the system can migrates from one minima to another.

Knowing all minima would enable to describe the relevant structures and conformations and their free energies, as well as the dynamics of structural transitions. Given a starting configuration, it is possible to find the nearest local minimum. "Nearest" in this context does not always imply "nearest" in a geometrical sense (i.e., the least sum of square

coordinate differences), but means the minimum that can be reached by systematically moving down the steepest local gradient.

There are three energy minimization methods. One, those that require only function evaluations. Examples are the simplex method and its variants. A step is made on the basis of the results of previous evaluations. Second those that use derivative information. Since the partial derivatives of the potential energy with respect to all coordinates are known in MD programs (these are equal to minus the forces) this class of methods is very suitable as modification of MD programs. Third those that use second derivative information as well. Gromacs focuses on the starting two methods for energy minimization.

1.3 MM-PBSA

In simulation of these solvated states the majority of the energy contributions would come from solvent-solvent interactions and the fluctuations in total energy would be an order of magnitude larger than binding energy and the calculation would take an inordinate amount of time to converge. That's why a more effective method is to divide up the calculation.

Molecular mechanics Poisson−Boltzmann surface area approach calculate and combines three energetic terms for getting the change in the free energy on binding. The first term corresponds to a change in the potential energy in the vacuum. It includes both bonded terms such as bond, angle, and torsion energies as well as non-bonded terms such as van der Waals and electrostatic interactions. The second term accounts for the desolvation of the different species. It is expressed by the sum of two energy terms which are the polar and nonpolar solvation energies using a solvation model. The third term accounts for the configurational entropy associated with complex formation in the gas phase. We need to calculate the binding free energy to find the relative binding stability.

For a protein-ligand complex the binding free energy can be expressed as:

$$\Delta G_{\text{binding}} = G_{\text{complex}} - (G_{\text{protein}} + G_{\text{ligand}}) \tag{1.3}$$

Here, G_{complex} is the total free energy of the protein−ligand complex and G_{protein} and G_{ligand} are total free energies of the isolated protein and ligand in solvent. The individual free energy can be calculated as:

$$G_x = \langle E_{MM} \rangle - TS + \langle G_{solvation} \rangle \quad (1.4)$$

Here, x is the protein or ligand or protein-ligand complex. $\langle E_{MM} \rangle$ is the average molecular mechanics potential energy in a vacuum. TS refers to the entropic contribution to the free energy in a vacuum. Vacuum potential energy is composed of:

$$E_{MM} = E_{bonded} + E_{nonbonded} = E_{bonded} + (E_{vdW} + E_{elec}) \quad (1.5)$$

The van der waal interaction is calculated using Lenard Jones potential and electronic using coulombic interaction. Free energy of solvation ($G_{solvation}$) is the energy required to transfer a solute from vacuum into the solvent. It is basically addition of Free energy of polar and nonpolar contribution to the solvation free energy.

$$G_{solvation} = G_{polar} + G_{nonpolar} \quad (1.6)$$

We calculated the polar contribution by solving Poisson−Boltzmann (PB) equation which is a second order partial differential equation given by

$$\nabla \cdot [\varepsilon(r) \nabla \cdot \varphi(r)] - \varepsilon(r) \kappa(r)^2 \sinh[\varphi(r)] + \frac{4\pi \rho^f(r)}{kT} = 0 \quad (1.7)$$

Where $\phi(r)$ is electrostatic potential, $\varepsilon(r)$ is the dielectric constant, and $\rho^f(r)$ is the fixed charge density. The term $\kappa(r)^2$ is related to the reciprocal of Debye length which is dependent on the ionic strength of the solution.

Nonpolar contribution includes both repulsive and attractive forces between solute and solvent that are generated by cavity formation and van der Waals interactions,

$$G_{nonpolar} = G_{cavity} + G_{vdW} \quad (1.8)$$

Where G_{cavity} is work done by the solute to create a cavity in the solvent and depends on the shape and geometry of the solute. G_{vdW} is the attractive van der Waals energy between solvent and solute. We estimate these terms using SASA-ONLY non-polar model.

Solvent accessible surface area (SASA) model is based on the assumption that the SASA is linearly dependent on the $G_{nonpolar}$ term and is calculated as

$$G_{\text{nonpolar}} = \gamma A + b \tag{1.9}$$

Here, γ is a coefficient related to surface tension of the solvent, A is SASA, and b is fitting parameter.[21,22]

1.4 Root mean square fluctuation

It is the measure of the distance between the atoms over a period of time with respect to reference frame.

$$\text{RMSF}_i = \left[\frac{1}{T} \sum_{t_j=1}^{T} |\mathbf{r}_i(t_j) - \mathbf{r}_i^{\text{ref}}|^2 \right]^{1/2} \tag{1.10}$$

1.5 Root mean square deviation

For RMSD the average is taken over the particles, giving time specific values with respect to reference frame

$$\text{RMSD}(t) = \left[\frac{1}{M} \sum_{i=1}^{N} m_i |\mathbf{r}_i(t) - \mathbf{r}_i^{\text{ref}}|^2 \right]^{1/2} \tag{1.11}$$

2: Methodology

2.1 Mutation

We have taken the PDBID: 4JVH mRNA bound QKI protein coordinate file for the mutation. We used x3dna v2.3.3 tool to make 5 mutations as:[23]

1. AAUAACAA
2. ACAAACAA
3. ACUCACAA
4. ACUACCAA
5. ACUAAAAA
6. ACUAACAA (Wild type)

2.2 Gromacs

For the coordinates we have taken PDBID: 4JVH file and use Gromacs Version 5.1.2 for the simulation. We used CHARMM27 force field and to solvate our molecule we used TIP3P water model to solvate the box. Na^+ and Cl^- ions were added to neutralise. The minimization was done for that system is restrained under maximum force 1000.0 kJ/mol/nm with 50000 steps using verlet cut-off scheme. Long-range electrostatic interactions were treated using the Particle Mesh Ewald (PME) technique. The structure obtained was then equilibrated by NVT and then NPT ensemble for 200ps. Using non-bonded verlet scheme and Particle Mesh Ewald for long-range electrostatics. Temperature coupling was done according V-rescale modified Berendsen thermostat taking 300K as reference temperature. For pressure coupling we used Parrinello-Rahman. After equilibration the system is then run for 5ns.

2.3 MMPBSA

For this we have used the open source tool g_mmpbsa where we have provided it the trajectory, index file and the coordinate file to calculate binding free energy.

A) For polar calculations:

Charge of positive ions pcharge = 1

Radius of positive charged ions prad = 0.95

Concentration of positive charged ions pconc = 0.150

Charge of negative ions ncharge = -1

Radius of negative charged ions nrad = 1.81

Concentration of negative charged ions nconc = 0.150

Solute dielectric constant pdie = 2

Solvent dielectric constant sdie = 80

Reference or vacuum dielectric constant vdie = 1

Solvent probe radius srad = 1.4

Method used to map biomolecular charges on grid. chgm = spl0 or spl2 or spl4

chgm = spl4

Model used to construct dielectric and ionic boundary. srfm = smol or spl2 or spl4

srfm = smol

Value for cubic spline window. Only used in case of srfm = spl2 or spl4. swin = 0.30

Number of grid point per Angstrom^2. Not used when (srad = 0.0) or (srfm = spl2 or spl4) sdens = 10

Temperature in K temp = 300

Type of boundary condition to solve PB equation. bcfl = zero or sdh or mdh or focus or map bcfl = mdh

Non-linear (npbe) or linear (lpbe) PB equation to solve PBsolver = lpbe

B) For non-polar SASA model:

Gamma (Surface Tension) kJ/(mol Angstrom^2) gamma = 0.0226778

Probe radius for SASA (A) sasrad = 1.4

Offset (c) kJ/mol sasaconst = 3.84982

2.4 Native contact code for within 5 angstrom pairs

```
proc residueContactPairs { cutoff sel1 sel2 } {
  set cl [ measure contacts $cutoff $sel1 $sel2 ]
  if { $cl == {} } {
    return {}
  }

  # List of (duplicated) residues in contact
  set l1 [lindex $cl 0]
  set l2 [lindex $cl 1]
  set n [llength $l1]

  # Now uniquify the pairs
  array set pairs {}
  for {set i 0} {$i < $n} {incr i} {
    set i1 [lindex $l1 $i]
    set i2 [lindex $l2 $i]
    set tmp [atomselect top "index $i1"]
    set r1 [$tmp get resid]
    $tmp delete
    set tmp [atomselect top "index $i2"]
    set r2 [$tmp get resid]
    $tmp delete

    #Remove self contacts
    if {$r1 !=$r2} {
      #count AB/BA contacts only once
      if { ! [info exists pairs($r2,$r1)] } {
        set pairs($r1,$r2) 1
      }
    }
 }

  set plist {}
  foreach p [array names pairs] {
    set pl [split $p ,]
    lappend plist $pl }
  # Return contact pairs
    return $plist
}


mol new md55pbc.xtc type xtc waitfor all #trajectory file
mol addfile npt.gro type gro molid 0
set outfile [open "AUAAC-RNAcontacts.dat" w]
set num_steps [molinfo top get numframes]
for {set frame 0} {$frame < $num_steps} {incr frame} {
 set sel1 [atomselect top "nucleic and noh" frame $frame]
  set sel2 [atomselect top "protein and noh" frame $frame]
  set Aclist [residueContactPairs 5.0 $sel1 $sel2]
  puts $outfile "$frame $Aclist"
  $sel1 delete
  $sel2 delete
}
close $outfile
exit
```

3: Result and discussion

3.1 Root mean square fluctuations

Using GROMACS we have calculated the RMSF values for C3’ atom of mRNA backbone with respect to energy minimized structure as reference.

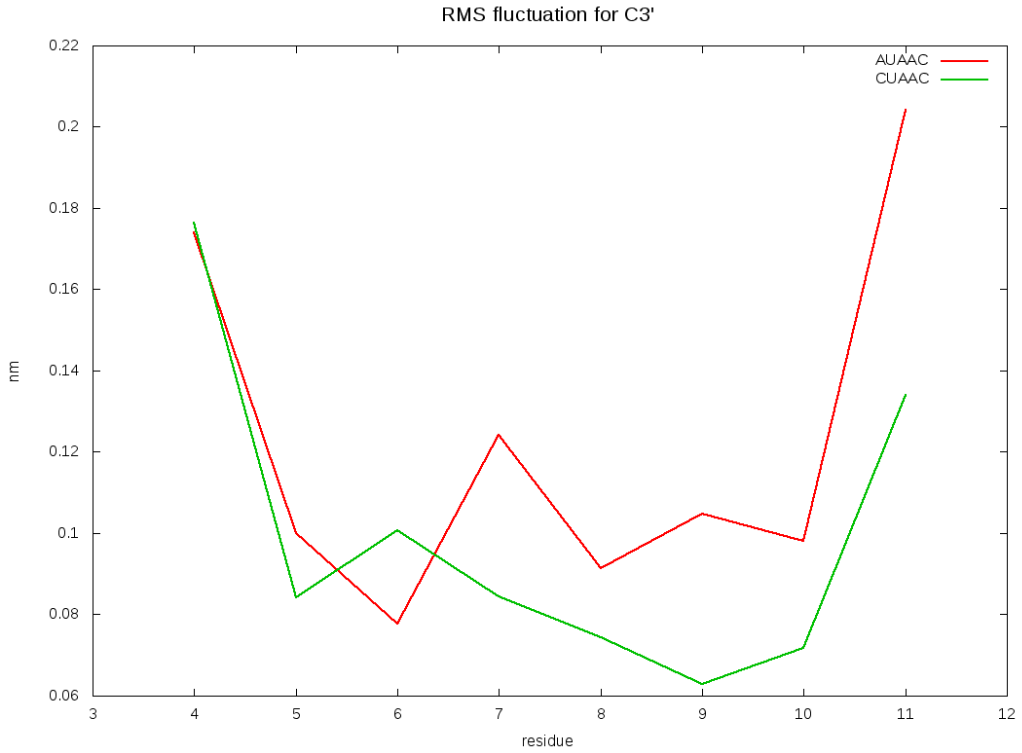

Fig. 3.1 RMSF for AUAAC vs CUAAC

Here the residue 5 is mutated with adenine with respect to wild type cytosine. There is slight increase in the fluctuation for the mutated adenine. And it keep on increasing for the rest of the mRNA residues too apart from residue 6 (uracil) with respect to their wild types.

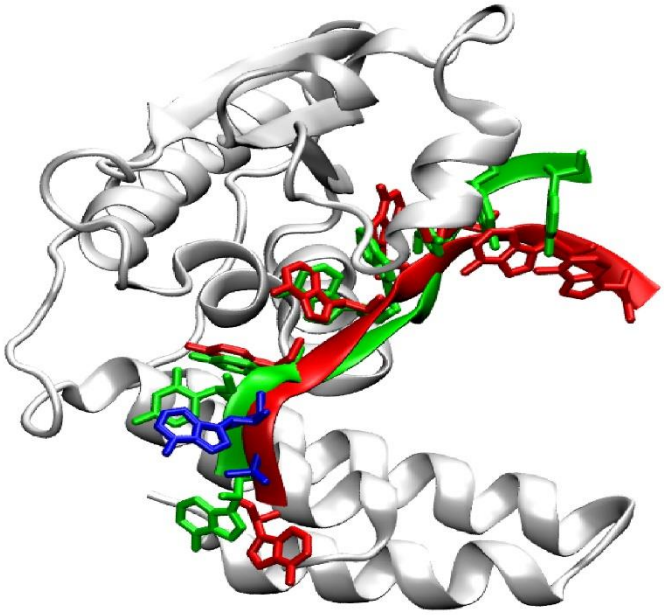

Fig 3.2 Visualization of AUAAC VS CUAAC crystal structure.

Red depicts the mutated mRNA and green wild type. Grey is the protein and blue shows the mutated adenine nucleotide.

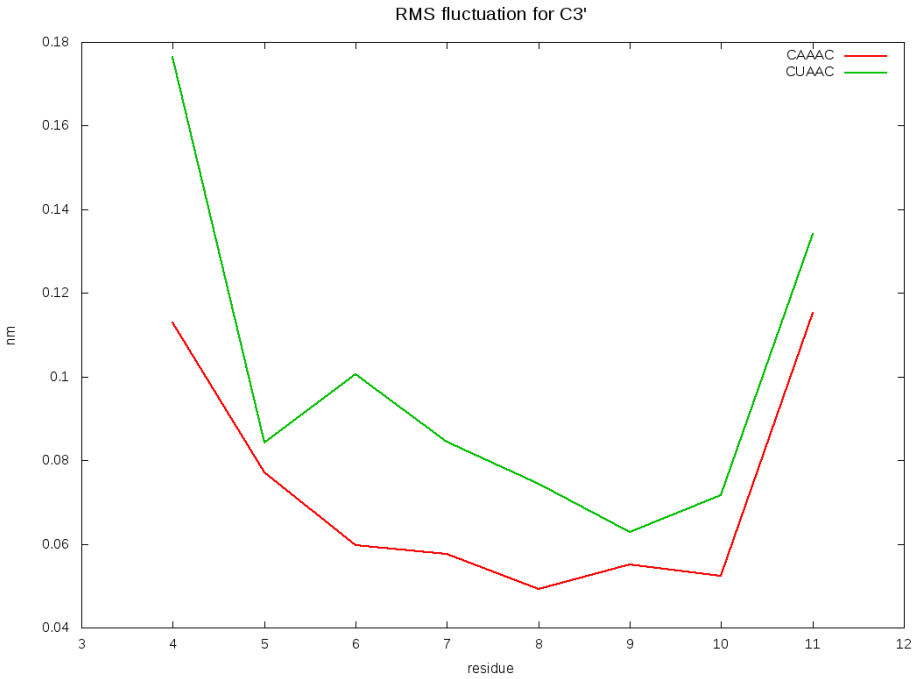

Fig. 3.3 RMSF for CAAAC vs CUAAC

Here the residue 6 is mutated with adenine with respect to wild type uracil. The fluctuation are less for the mutated adenine. And the overall trend shows decrement from wild type.

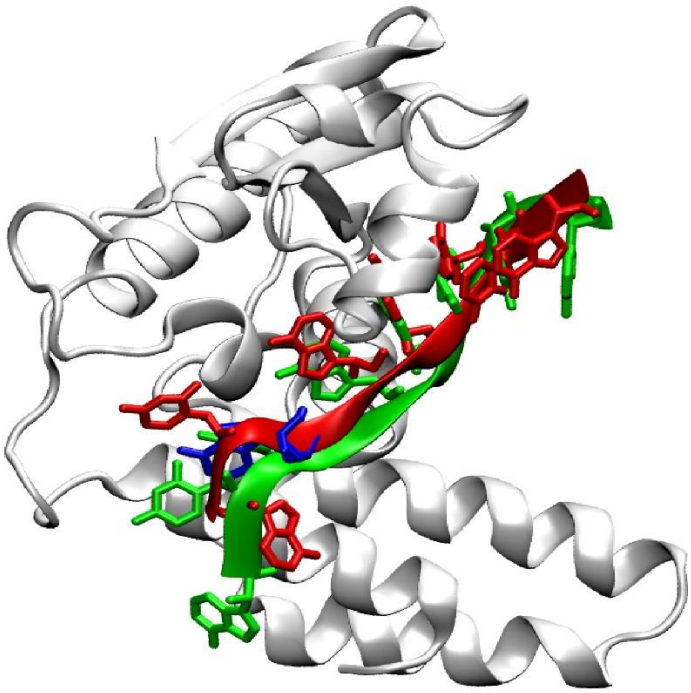

Fig 3.4 Visualization of CAAAC VS CUAAC crystal structure

Red depicts the mutated mRNA and green wild type. Grey is the protein and blue shows the mutated adenine nucleotide.

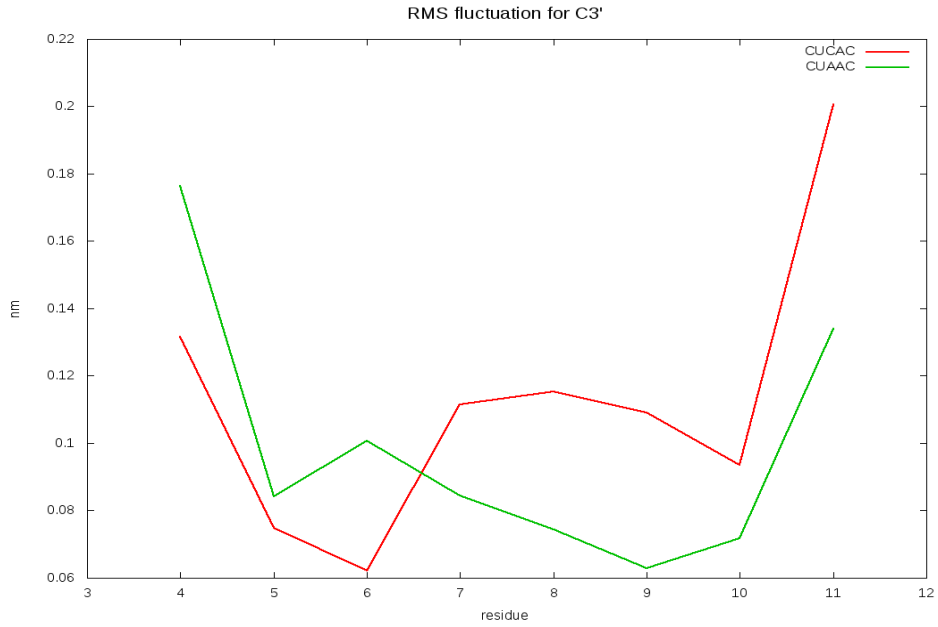

Fig. 3.5 RMSF for CUCAC vs CUAAC

Here the residue 7 is the mutated with cytosine with respect to wild type adenine. Mutated cytosine shows an increase in fluctuation with respect to wild type adenine.

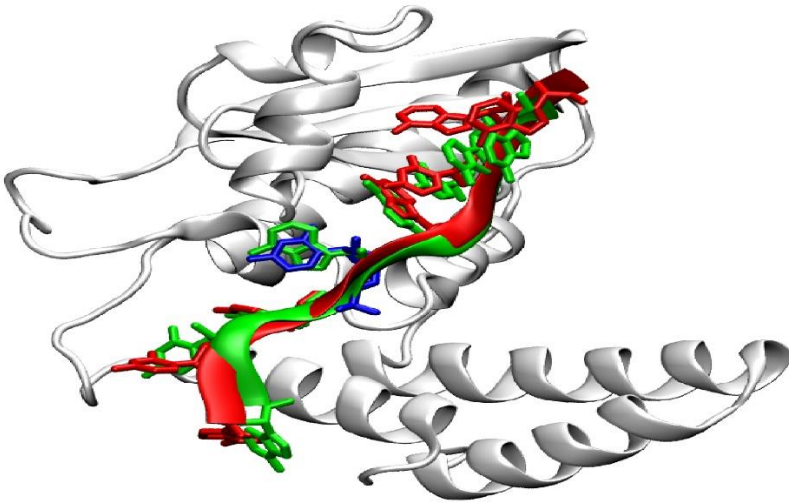

Fig 3.6 Visualization CUCAC VS CUAAC crystal structure

Red depicts the mutated mRNA and green wild type. Grey is the protein and blue shows the mutated cytosine nucleotide.

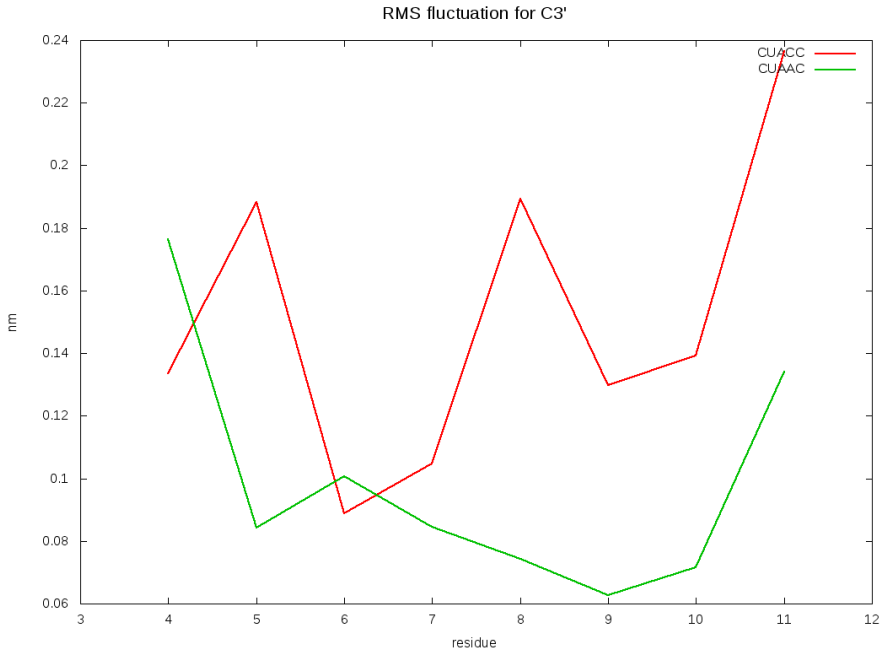

Fig. 3.7 RMSF for CUACC vs CUAAC

Here the residue 8 is the mutated with cytosine with respect to wild type adenine. Mutated cytosine shows an increase in fluctuation with respect to wild type adenine.

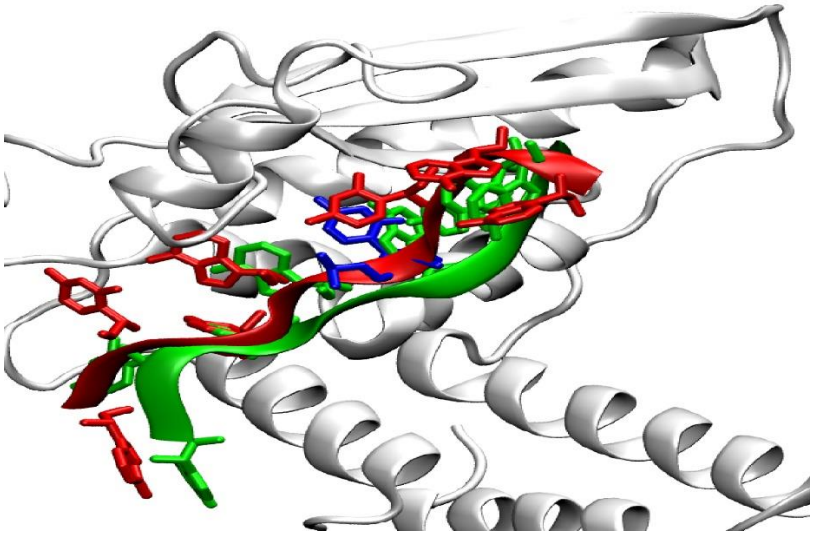

Fig 3.8 Visualization CUACC VS CUAAC crystal structure

Red depicts the mutated mRNA and green wild type. Grey is the protein and blue shows the mutated cytosine nucleotide.

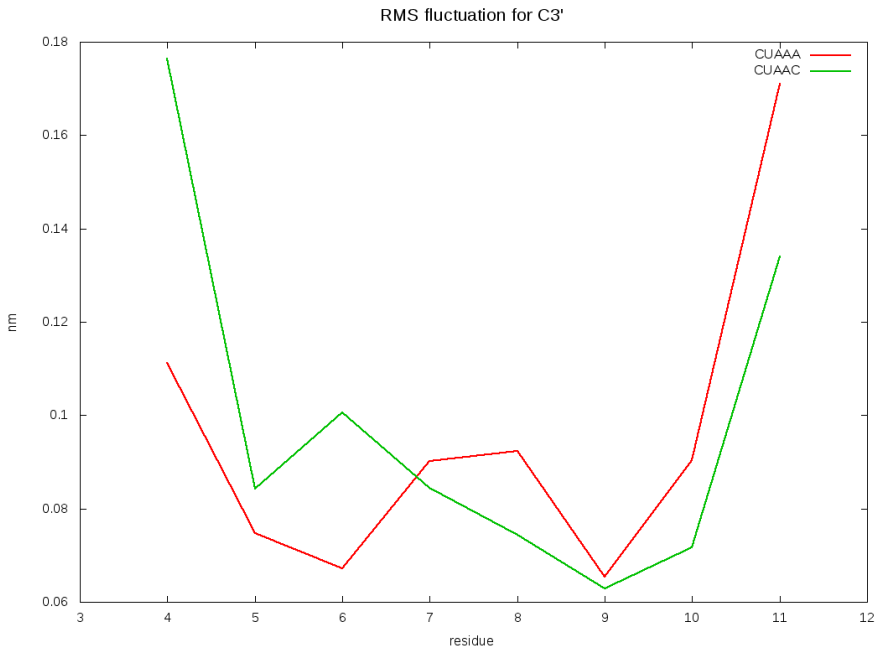

Fig. 3.9 RMSF for CUAAA vs CUAAC

Here the residue 9 is the mutated with adenine with respect to wild type cytosine. The mutated adenine shows an increase in fluctuation with respect to wild type cytosine.

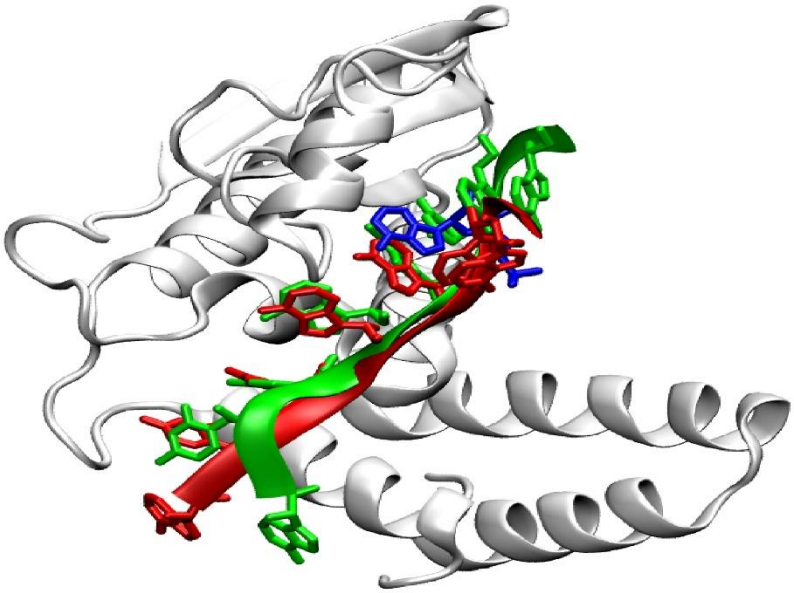

Fig3.10 Visualization CUAAA VS CUAAC crystal structure

Red depicts the mutated mRNA and green wild type. Grey is the protein and blue shows the mutated adenine nucleotide.

3.2 Root mean square deviation

Using VMD[24] RMSD trajectory tool we have calculated it for nucleic backbone with respect to equilibrated structure.

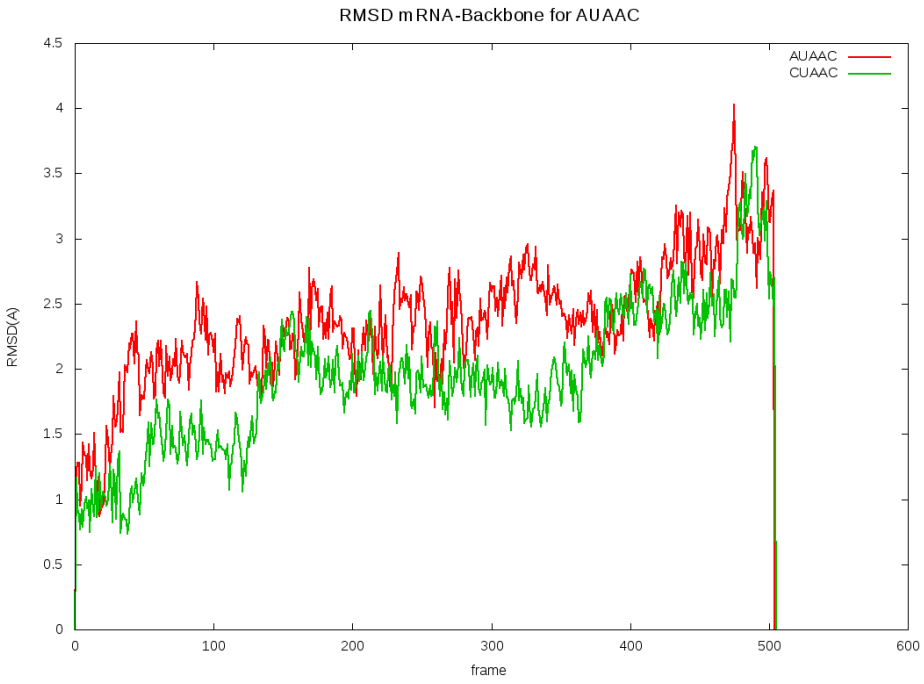

Fig. 3.11 RMSD for AUAAC Vs CUAAC

For most of the run rmsd values for the mutated mRNA were more than wild type.

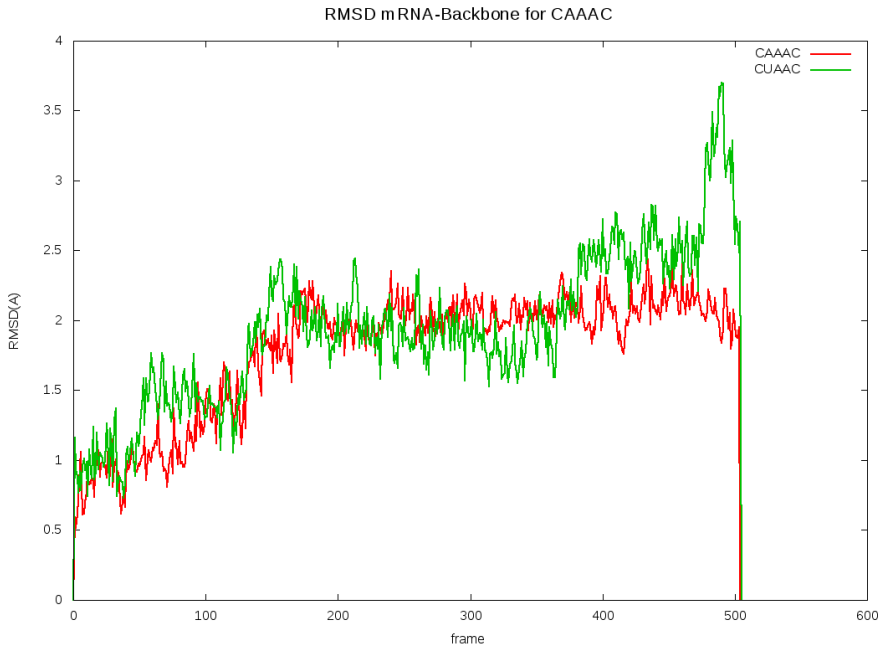

Fig. 3.12 RMSD for CAAAC Vs CUAAC

The rmsd values were less for mutated in comparison to wild type for most of the run.

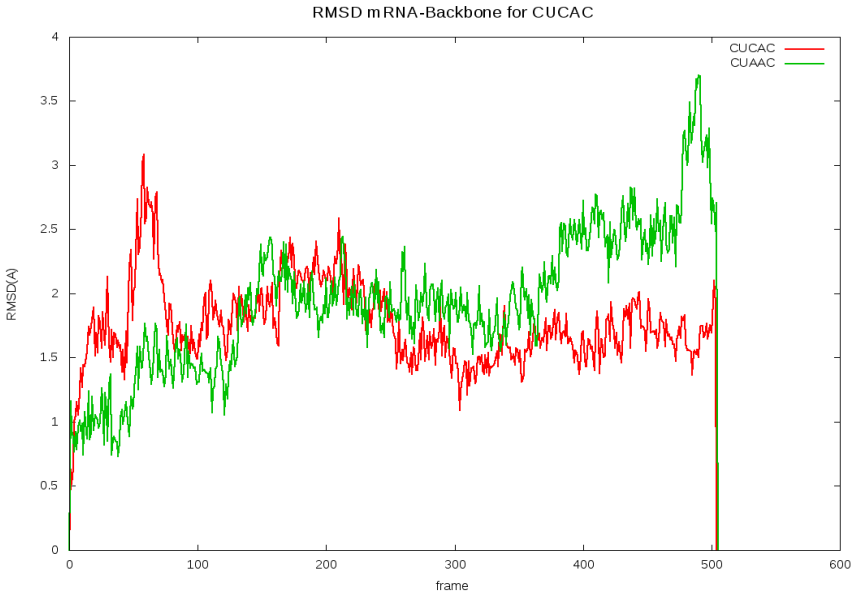

Fig. 3.13 RMSD for CUCAC Vs CUAAC

At intial run the rmsd values were more than wild type but after 250 frames we saw decrement in it.

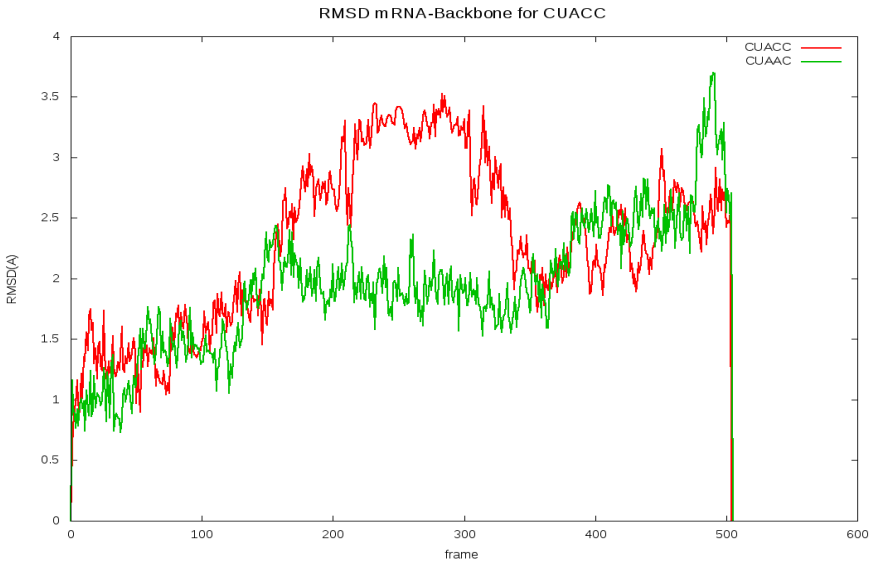

Fig. 3.14 RMSD for CUACC Vs CUAAC

At initial run rmsd was almost same for both but around 200 to 350 there is a sudden jump in rmsd for mutated one and then it get backs along with wild type and then ending with a steep rise.

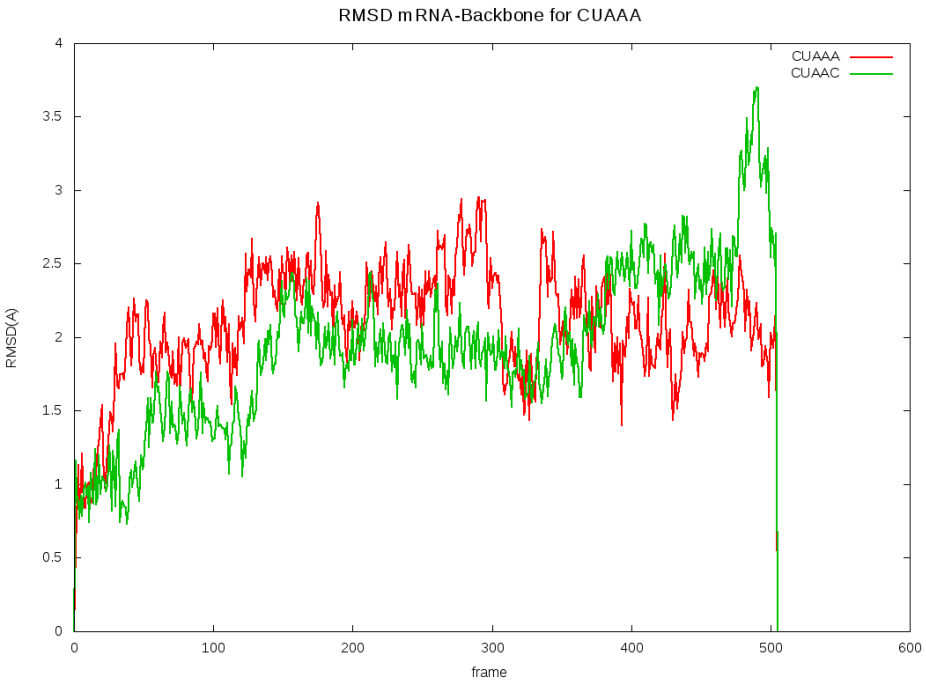

Fig. 3.15 RMSD for CUAAA Vs CUAAC

Till the mid stage of the run rmsd values for mutated were more than wild type but after 400 frames we see it goes below the wild type.

3.3 Binding Energy (kJ/mol)

To calculate the binding energy between QKI STAR protein and mRNA we have used MMPBSA technique.

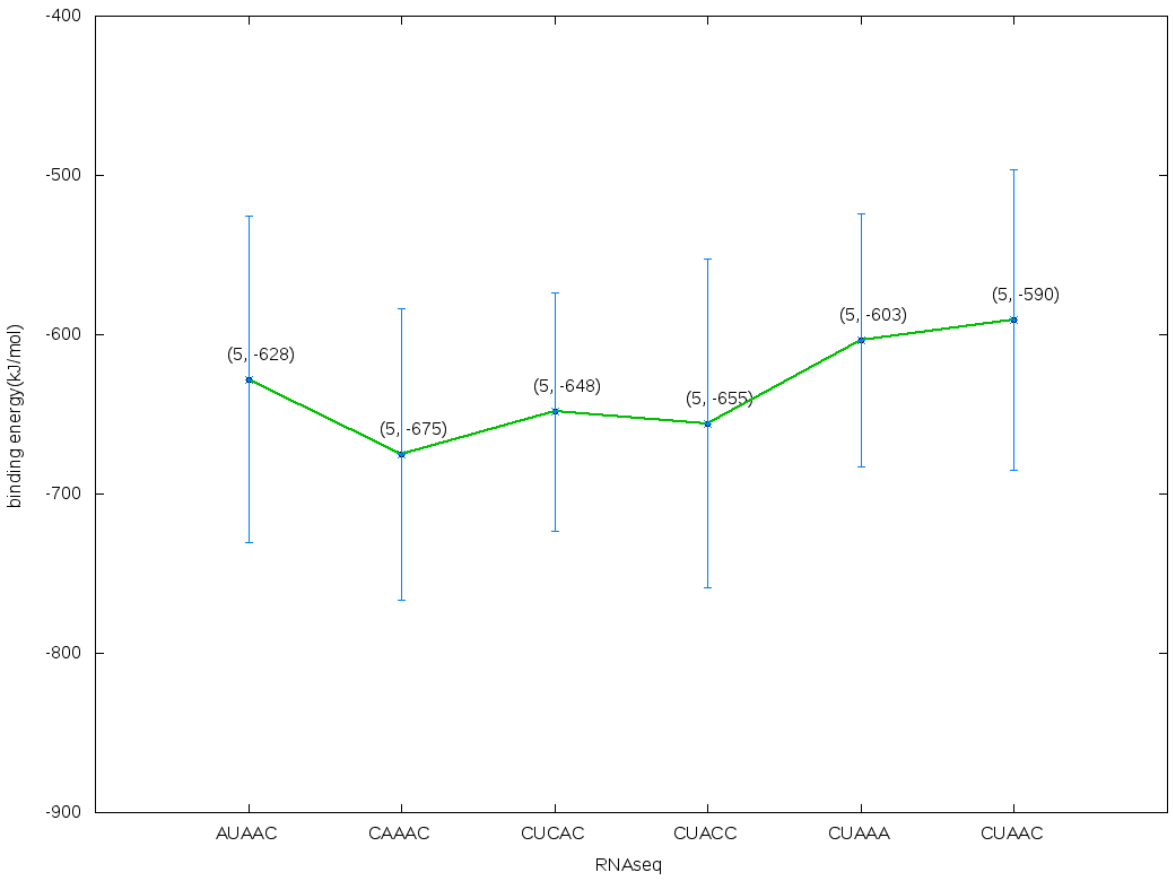

Fig 3.16 MMPBSA Binding Energy for different residues vs wild type

It shows that the CAAAC mutation is the most stable one for 5ns run.

3.4 Close Contacts

Using the trajectory file and coordinate gro file we have calculated the close contacts within 5 angstrom for 5ns run for each mutation. Here x axis represent the protein residues and y axis mRNA. The numbers are showing the percentage the pair was within 5 angstrom for the whole run. The dark spaces shows 0 percent.

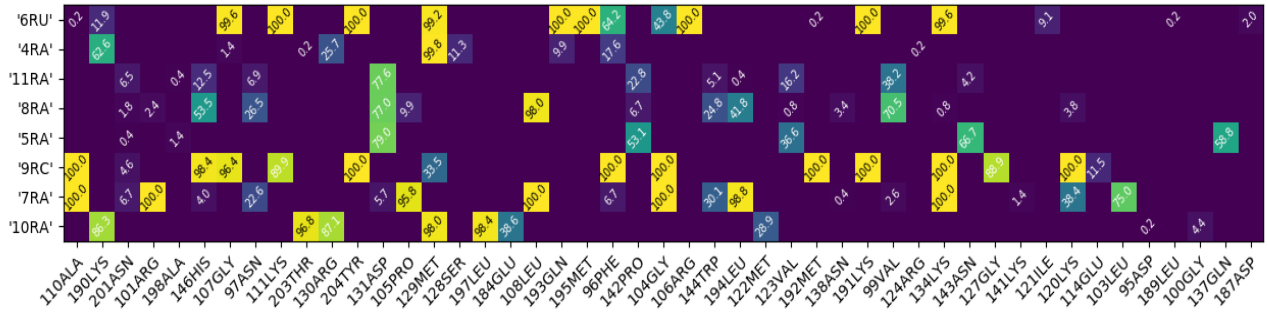

Fig. 3.17 Contact pair percentage for AUAAC for 5ns run

Wild type cytosine shows relatively more percent close contact pairs than mutated adenine.

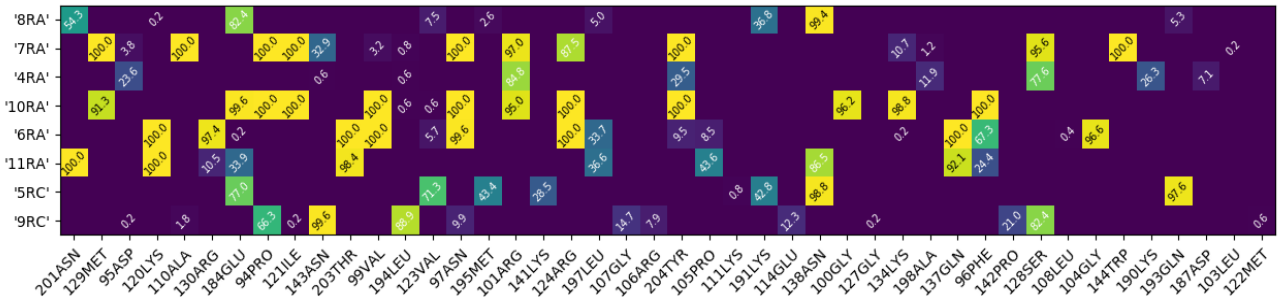

Fig. 3.18 Contact pair percentage for CAAAC for 5ns run

Although the wild type uracil shows more no. of contact pair but mutated adenine shows higher close contact percent than uracil in the whole run.

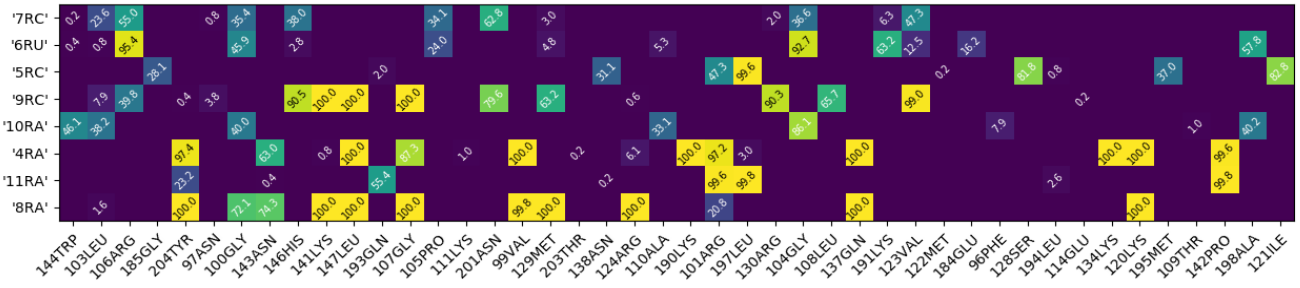

Fig. 3.19 Contact pair percentage for CUCAC for 5ns run

Both the wild type adenine and mutated cytosine shows not a single 90+ % contact with any protein residue.

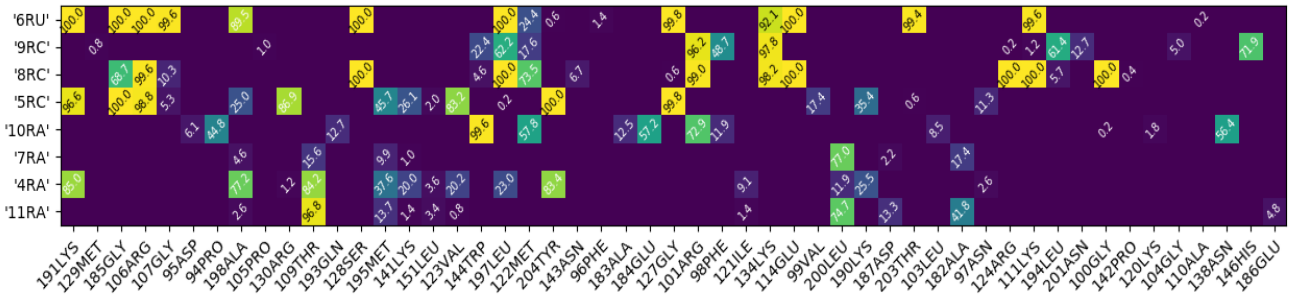

Fig. 3.20 Contact pair percentage for CUACC for 5ns run

The mutated cytosine shows more contact with protein residue than wild type adenine.

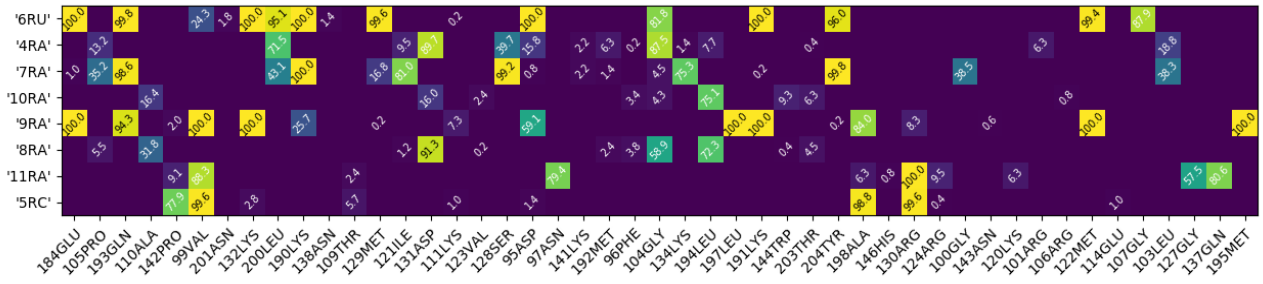

Fig. 3.21 Contact pair percentage for CUAAA for 5ns run

The mutated adenine shows more contact with protein residues than its wild type cytosine.

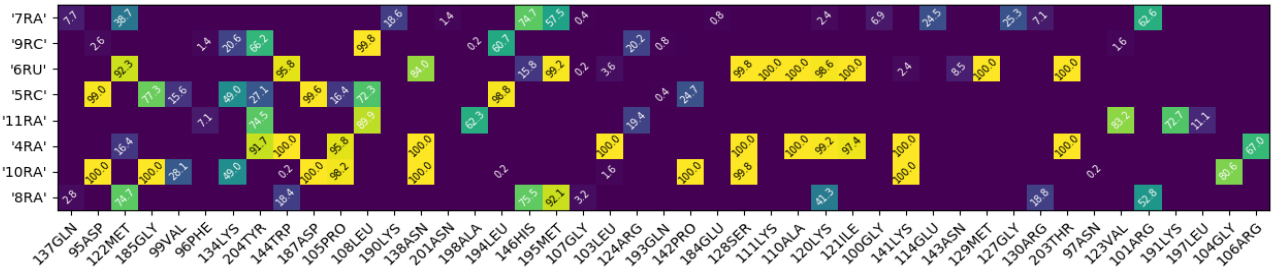

Fig. 3.22 Contact pair percentage for CUAAC for 5ns run

Resid 6 uracil residue is closest to most number of close proteins for the whole run

Bibliography

1. Vernet, C. & Artzt, K. STAR, a gene family involved in signal transduction and activation of RNA. Trends Genet. 13, 479–484 (1997).
2. Chénard, C. A. & Richard, S. New implications for the QUAKING RNA binding protein in human disease. J. Neurosci. Res. 86, 233–242 (2008).
3. Teplova, M. et al. Structure-function studies of STAR family Quaking proteins bound to their in vivo RNA target sites. Genes Dev.27, 928–940 (2013).
4. Galarneau, A. & Richard, S. Target RNA motif and target mRNAs of the Quaking STAR protein. Nat. Struct. Mol. Biol. 12, 691–698 (2005).
5. Hafner, M. et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 141,129–141 (2010).
6. Jungkamp, A.-C. et al. In vivo and transcriptome-wide identification of RNA binding protein target sites. Mol. Cell 44, 828–840(2011).
7. Chen T, Richard S 1998. Structure-function analysis of Qk1: A lethal point mutation in mouse quaking prevents homodimerization. Mol Cell Biol 18: 4863–4871
8. Chen T, Damaj BB, Herrera C, Lasko P, Richard S 1997. Self-association of the single-KH-domain family members Sam68, GRP33, GLD-1, and Qk1: Role of the KH domain. Mol Cell Biol 17: 5707–5718
9. Ryder SP, Williamson JR 2004. Specificity of the STAR/GSG domain protein Qk1: Implications for the regulation of myelination. RNA 10: 1449–1458
10. Bockbrader, K. & Feng, Y. Essential function, sophisticated regulation and pathological impact of the selective RNA-binding protein QKI in CNS myelin development. Future Neurol. 3, 655–668 (2008).
11. Arning, S., Grüter, P., Bilbe, G. & Krämer, A. Mammalian splicing factor SF1 is encoded by variant cDNAs and binds to RNA. RNA 2, 794–810 (1996).
12. Liu, Z. et al. Structural basis for recognition of the intron branch site RNA by splicing factor 1. Science 294, 1098–1102 (2001).
13. Matter, N., Herrlich, P. & König, H. Signal-dependent regulation of splicing via phosphorylation of Sam68. Nature 420, 691–695(2002).

14. Francis, R., Barton, M. K., Kimble, J. & Schedl, T. gld-1, a tumor suppressor gene required for oocyte development in Caenorhabditis elegans. Genetics 139, 579–606 (1995).
15. Lee, M.-H. & Schedl, T. Translation repression by GLD-1 protects its mRNA targets from nonsense-mediated mRNA decay in C. elegans. Genes Dev. 18, 1047–1059 (2004).
16. Schumacher, B. et al. Translational repression of C. elegans p53 by GLD-1 regulates DNA damage-induced apoptosis. Cell 120,357–368 (2005).
17. Crittenden, S. L. et al. A conserved RNA-binding protein controls germline stem cells in Caenorhabditis elegans. Nature 417,660–663 (2002).
18. Francis, R., Maine, E. & Schedl, T. Analysis of the multiple roles of gld-1 in germline development: interactions with the sex determination cascade and the glp-1 signaling pathway. Genetics 139, 607–630 (1995).
19. Jan, E., Motzny, C. K., Graves, L. E. & Goodwin, E. B. The STAR protein, GLD-1, is a translational regulator of sexual identity in Caenorhabditis elegans. EMBO J. 18, 258–269 (1999).
20. Jones, A. R. & Schedl, T. Mutations in gld-1, a female germ cell-specific tumor suppressor gene in Caenorhabditis elegans, affect a conserved domain also found in Src-associated protein Sam68. Genes Dev. 9, 1491–1504 (1995).
21. Kumari et al (2014) g_mmpbsa - A GROMACS tool for high-throughput MM-PBSA calculations. J. Chem. Inf. Model. 54:1951-1962
22. Baker et al. (2001) Electrostatics of nanosystems: Application to microtubules and the ribosome. Proc. Natl. Acad. Sci. USA 98:10037-10041.
23. Andrew Colasanti, Xiang-Jun Lu & Wilma K. Olson (2013) Analyzing and building nucleic acid structures with 3DNA. Journal of visualized experiments. JoVE, 74, e4401
24. Humphrey, W., Dalke, A. and Schulten, K., "VMD - Visual Molecular Dynamics", J. Molec. Graphics, 1996, vol. 14, pp. 33-38.