Analysis of DGOR protein docked with ligand D-galactonate

Kapil Yadav

MS14037

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

Indian Institute of Science Education and Research Mohali

December 2019

Certificate of Examination

This to certify that the dissertation titled **“Analysis of DGOR protein docked with ligands D-galactonate ”** submitted by **Mr. Kapil Yadav**(Reg. No. **MS14037**) 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. Debrina Jana Dr. Rachna Chaba Dr. Monika Sharma

(Supervisor)

Dated: December 5, 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 the bibliography.

Kapil Yadav
(Candidate)

Dated: December 5, 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 thank all those who have helped and supported me directly or indirectly throughout my BS-MS time. Without them, this thesis would not have been possible.

I would like to thank all the people who helped me during my project work especially my project guide Dr Monika Sharma who gave me the opportunity to work with her and made me keen to this topic.

I am most thankful to my parents who supported and motivated me throughout my life.

My life in IISER would be different if I have not met my Jugaadis which includes Akash, Ankit, Ajay, Mohit, Munish, Pranshu, Ravi, Ravineet, Sahil and Vishal who always backed me during my whole degree time and helped me. I would also like to thank Dr. Gopal and Dr. Pankaj Dubey.

I would also like to thank my project partner Mohit Kumar with whom I work in collaboration with.

List of figures:

Abstract

D-galactonate is long known as a source of sugar acid for E.coli family. It was recently found that DgoR deletion accelerates the growth of E. coli in D-galactonate concomitant with the strong constitutive expression of dgo genes(10). We are going to discuss the effect of D-galactonate as a ligand on the binding property of the protein DgoR (strain K=12). In order to do that we are going to compare two forms of DgoR protein by running molecular simulation on the system consisting of DgoR and DNA.

Contents

1. Introduction

1.1 Transcriptional Regulators:-

All cells contain a set of genes, which can be thought of as a set of instructions for making each of a very large number of proteins . The creation of a protein from its gene is called gene expression. Transcription is the term given for the process of conversion of DNA to RNA. Transcriptional regulation is referred to the means by which a cell regulates the process of transcription. Transcriptional regulation is a subject of interest for decades. The centre of this study lies in the selection of transcription start site(TSS), defined as the first transcribed genome nucleotide of a transcript(1). The process of this transcription start site selection is influenced by the presence of general transcription factors (GTFs) and their DNA binding activities(1). A single gene can be regulated in an infinite number of possible ways, from altering the number of copies of RNA that are transcribed, to the temporal control of when the gene is transcribed. A substance such as a protein or an enzyme that contribute to the cause of a specific biochemical reaction or bodily process is called a transcriptional factor. There are two types of transcriptional regulators, coactivator and corepressor. A protein that works with transcription factors to increase the rate of gene transcription is called a coactivator while protein that works with transcription factors to decrease the rate of gene transcription is called a corepressor. Classical definitions of activating regulatory elements are focused on two classes: promoters and enhancers, where the first category defines where transcription is initiated, and the other, elements that amplify such transcription initiation(1). Understanding the regulation of transcription is by nature far more complex in eukaryotic cells than in prokaryotic cells. Prokaryotes and eukaryotes shows basic similarities in gene transcription, one of which is that the RNA polymerase binds upstream of its gene on the promoter to initiate the process of transcription(2). Multicellular eukaryotes controls the process of transcriptional regulation through more complex and spatial regulation of gene expression(2).

1.2 DgoR:

DgoR is a FadR subfamily within GntR family transcriptional regulator(3). Human urinary tract bacteria of E. coli is introduced in the mouse gut and it has shown missense mutations in DgoR. The common gut bacterium, Escherichia coli, can utilize a variety of sugar acids, i.e., hexonates, hexuronates, hexuronides and aldarates, as carbon and energy source (3). DNA-binding operator sequence of DgoR is [5'-TTGTA(G/C)TACA(A/T)-3'](3). DgoR (with a total of 230 residues) sequence generated from FASTA is(3):

(OS=Escherichia coli ,strain K12, GN=dgoR)

MTLNKTDRIVITLGKQIVHGKYVPGSPLPAEAELCEEFATSRNIIREVFRSLMAKRLIE
MKRYRGAFVAPRNQWNYLDTDVLQWVLENDYDPRLISAMSEVRNLVEPAIARWAA
ERATSSDLAQIESALNEMIANNQDREAFNEADIRYHEAVLQSVHNPVLQQLSIAISSLQ
RAVFERTWMGDEANMPQTLQEHKALFDAIRHQDGDAAEQAALTMIASSTRRLKEIT

(1.1.1)

1.3 D-galactonate:

D-galactonate is a galactonate compound having D-configuration (dextrorotatory optical isomer). A dextrorotatory compound is a compound that rotates the plane of polarized light clockwise as it approaches the observer (to the right), If a compound is dextrorotatory, its mirror image counterpart is levorotatory. That is, it rotates the plane of polarized light counterclockwise (to the left). Formula of D-galactonate is: $C_6H_{12}O_7$ and molar mass of: 196.1553 g/mol(4). The whole family of galactonic acid has been manually annotated by the ChEBI team. According to them it belongs to the class of human metabolite (Any mammalian metabolite produced during a metabolic reaction in humans (*Homo sapiens*)).

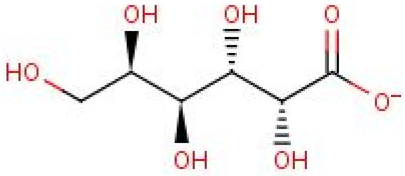


Figure1.1 (D-galactonate)

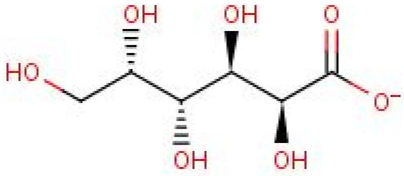


Figure1.2 (L-galactonate)

D-galactonate is a widely prevalent aldonic sugar acid. In Spite being suggested many times about the significance of D-galactonate's metabolism in the interaction of bacteria with their hosts, there are no records of its regulation in *E.coli* (DgoR)(3). Because of this utilization of these sugar acid, E.Coli has shown exponential growth in mammalian gut which resulted in the colonization of E.Coli bacteria. The sugar-acid metabolism are controlled by the specific transcriptional regulators whose DNA-binding properties are affected by binding to effectors, which could either be the sugar-acid itself or it's catabolic intermediate. D-galactonate, a hexanoate sugar acid, was first reported as a carbon source for E. coli in studies conducted in the 1970's (3). Through classical mutagenesis and biochemical approaches, it was shown that E. coli metabolizes D-galactonate through a modified form of the Entner-Doudoroff pathway by a set of genes arranged in a putative D-galactonate operon (3,5).

We humans also produce D-galactonate its production in patients suffering from galactosemia, who lack the enzyme of D-galactose metabolism. Several studies suggest the abundance occurrence of D-galactonate in nature itself(6). The just mentioned examples of D-galactonate presence in nature and human body arises the situation of investigating D-galactonate metabolism in enteric bacteria. This invokes the need of understanding the effects of D-galactonate on DgoR-DNA binding and their interaction.

1.4 Computer Simulations:

Molecular dynamics (MD) is a computer simulation with atoms and/or molecules interacting using some basic laws of physics. Computer simulations act as a bridge between microscopic length and timescales and the macroscopic world of the laboratory, that is, we provide some basic information about the interaction of the protein and receive information about a bulk of conformations of protein(7). A classical system uses molecular dynamics to calculate the balance and transportation properties(7). The very basic example of the application of MD simulation is in solving Newton's equations of motion for a system of N interacting atoms. In this particular case, we are more interested in observing the time-dependent behaviour of our molecular system. This provides us with detailed information on fluctuation and conformational changes of the proteins and nucleic acids, which is used to investigate the dynamics, structure and thermodynamics of the biological molecular system. Protein stability, conformational changes, and protein folding are the major application of molecular simulation dynamics that we will be working on.

An experiment is usually made on a macroscopic sample that contains an extremely large number of atoms or molecules sampling an enormous number of conformations. In statistical mechanics, averages corresponding to experimental observables are defined in terms of ensemble averages(7). An ensemble average is an average taken over a large number of replicas of the system considered simultaneously(7).

The ensemble average is given by

$$\langle A \rangle_{ensemble} = \iint dp^N dr^N A\left(p^N, r^N\right) \rho\left(p^N, r^N\right) \qquad (1.2.1)$$

where

$$A\left(p^N, r^N\right)$$

is the observable of interest and it is expressed as a function of the momenta, *p*, and the positions, *r*, of the system. The integration is over all possible variables of *r* and *p.*

The probability density of the ensemble is given by:-

$$\rho\left(p^N, r^N\right) = \frac{1}{Q}\exp\left[-H\left(p^N, r^N\right) / k_B T\right]$$

(1.2.2)

1.4.1 Molecular dynamics simulation:

This technique is basically a simple numerical integration of Newton's equations of motion. The particles of the system are allowed to move freely without any constraints and adjustments, according to the interactions happens between them. It is basically a time evolution system of the protein. One challenge of this technique is to calculate large number of particle interactions, for that small-time step will be required, so our simulations are restricted to small time scales like several picoseconds, nanoseconds.

1.4.2 Force fields and software used while modelling:

We are using Gromacs, AMBER, VMD, CatDCD, ParmEd while modelling and simulating DgoR.

GROningen MAchine for Chemical Simulations (GROMACS) is a molecular dynamics package mainly designed for simulations of proteins, lipids, and nucleic acids. GROMACS is very fast due to algorithmic and processor-specific optimization, typically running 3-10 times faster than many simulation programs(8). CHARMM (Chemistry at HARvard Macromolecular Mechanics) is a both a set of force fields and a software for molecular dynamics simulations and analysis. Includes united atom (CHARMM19) and all atom (CHARMM22, CHARMM27, CHARMM36) force fields(9). The CHARMM27 force field has been imported to GROMACS. AMBER (Assisted Model Building and Energy Refinement) refers both to a set of molecular mechanical force fields (AMBER94, AMBER96, AMBER99, AMBER99SB, AMBER99SB-ILDN, AMBER03, AMBERGS) for the simulation of biomolecules and a software for molecular simulation programs(10). CatDCD is built using the molfile reader/writer plugins as the basis for it's input file. it is used to concatenate two trajectory files or to convert the trajectory file from the format of one force field to another. ParmEd is a general tool for aiding in investigations of biomolecular systems using popular molecular simulation packages, like Amber, CHARMM, and OpenMM written in Python(11). It is used for a variety of modelling purposes such as manipulating the topology of the system (i.e., the *atoms*, *bonds*, valence angles, etc..). ParmEd is used for reading and writing of a wide range of different file formats (such as CHARMM PSF, parameter, topology, and coordinate files, Amber topology and coordinate files) (11).

2. Methodology:

2.1 Modelling of DgoR:

Two final models (two different PDB files) for the protein are made. One is APO (which do not contain D-galactonate, system only consists of DNA and protein), another one is HOLO (which contains: DNA, protein and the ligand D-galactonate). To make the structure file we take the protein sequence shown in eq. (1.1.1), for that protein sequence we do the psi-Blast search hits across the protein data bank(PDB) using NCBI online blast hit search server. The Blast search shows many possible proteins that show similarity with the given protein sequence. The most appropriate similar structure to model the given DgoR sequence is of FADR-DNA complex: transcriptional control of fatty acid metabolism in E.Coli.

After that, we used Modeller software to model DgoR (100 models) using the structure of FADR-DNA complex(12) downloaded from the protein data bank. Modeller is a software used for homology or comparative modelling of protein three-dimensional structures. We provided an alignment of the above given sequence (1.1.1) to be modelled with known related structure(i.e. FADR-DNA complex) and modeller automatically calculates a model containing all non-hydrogen atoms.

Sequence of the provided PDB file of the templette protein(12):

>1HW2:A|PDBID|CHAIN|SEQUENCE

MVIKAQSPAGFAEEYIIESIWNNRFPPGTILPAERELSELIGVTRTTLREVLQRLARDG
WLTIQHGKPTKVNNFWETSGNILETLARLDHESVPQLIDNLLSVRTNISTIFIRTAFRQ
HPDKAQEVLATANEVADHADAFAELDYNIFRGLAFASGNPIYGLILNGMKGLYTRIG
RHYFANPEARSLALGFYHKLSALCSEGAHDQVYETVRRYGHESGEIWHRMQKNLPG
DLAIQGR

(2.1.1)

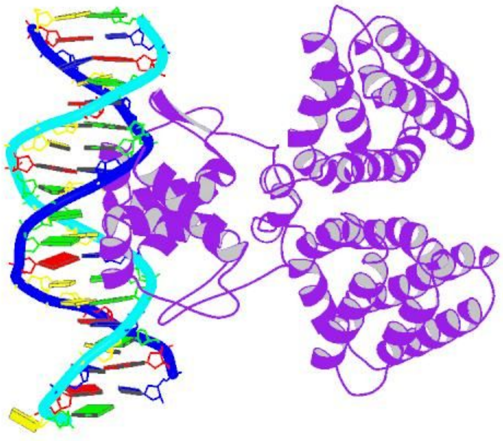


Figure 2.1 (FADR-DNA complex)

We use the following program (2.1.1) to make 100 models of DgoR and then we select the one with the best DOPE score.

```
from modeller import *

from modeller.automodel import *

env = environ()

a = automodel(env, alnfile='DGOR-1hw2.ali',

              knowns='1hw2', sequence='DGOR',

              assess_methods=(assess.DOPE,assess.GA341))

a.starting_model = 1

a.ending_model =100

a.make()
```

(2.1.1)

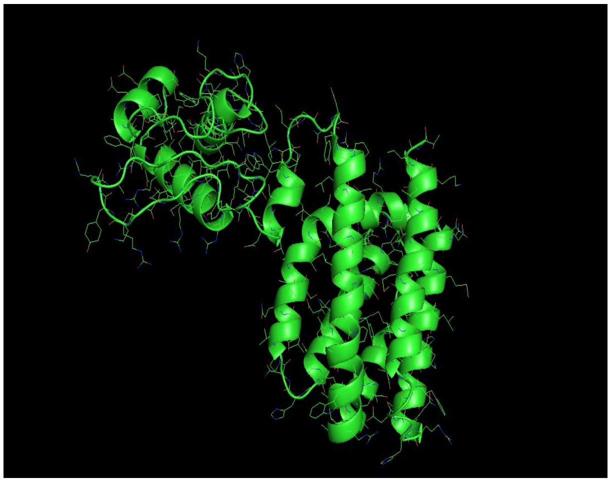
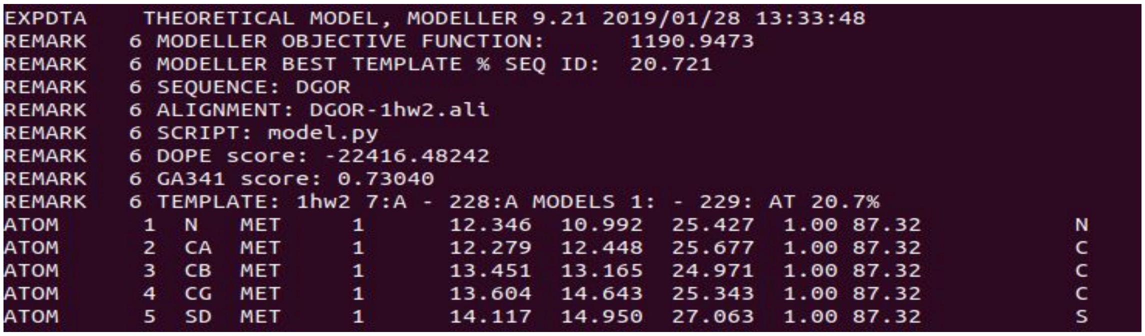


Figure 2.2 (Modelled DgoR)

2.2 PDB File:

The term PDB is used to express the Protein Data Bank (http://www.rcsb.org/pdb/)(13). A PDB file is a file containing the atomic coordinates of a molecule and other details which explains the structure of the biological system. Documents in the PDB incorporate data, for example, the name of the compound, the species and tissue from which is was gotten, creation, amendment history, diary reference, references, amino acid arrangement, stoichiometry, auxiliary structure areas, crystal lattice and symmetry group, lastly the ATOM and HETATM records containing the directions of the protein and any waters, particles, or different heterogeneous atom in the crystal. X-ray crystallography, NMR spectroscopy and cryo-electron microscopy are some techniques used to determine the location of each atom relative to another in the molecule which is further used to define the atomic coordinates in biological structure(13). Because of the limit of x-beam crystallography and NMR structure examination, the directions of hydrogen molecules are excluded in the PDB(13).



```
EXPDTA    THEORETICAL MODEL, MODELLER 9.21 2019/01/28 13:33:48
REMARK   6 MODELLER OBJECTIVE FUNCTION:      1190.9473
REMARK   6 MODELLER BEST TEMPLATE % SEQ ID:  20.721
REMARK   6 SEQUENCE: DGOR
REMARK   6 ALIGNMENT: DGOR-1hw2.ali
REMARK   6 SCRIPT: model.py
REMARK   6 DOPE score: -22416.48242
REMARK   6 GA341 score: 0.73040
REMARK   6 TEMPLATE: 1hw2 7:A - 228:A MODELS 1: - 229: AT 20.7%
ATOM      1  N   MET     1      12.346  10.992  25.427  1.00 87.32           N
ATOM      2  CA  MET     1      12.279  12.448  25.677  1.00 87.32           C
ATOM      3  CB  MET     1      13.451  13.165  24.971  1.00 87.32           C
ATOM      4  CG  MET     1      13.604  14.643  25.343  1.00 87.32           C
ATOM      5  SD  MET     1      14.117  14.950  27.063  1.00 87.32           S
```

Figure 2.3 (PDB file of DgoR)

2.3 Dimer of DgoR and it’s docking with D-galactonate and DNA:

PyMOL is a open-source software available for model visualization and used in structural biology(14). The *Py* part of the software's name refers that the programming language is Python(14).

Pymol adds crystallographic symmetry to the protein, if the protein is capable of forming the dimer i.e. if it has a symmetry then we can generate the dimer of protein by using the “align” command of the pymol(14). “align” performs a sequence alignment followed by a structural superposition, and then carries out zero or more cycles of refinement in order to reject outliers, the program then returns the RMSD value for all the aligned atoms. If one wants to perform an sequence-independent structure based alignment then they can use “super” command of pymol. It aligns two selections on the basis of their structures(14). super is more preferred than align for proteins with low sequence similarity.

After the formation of the dimer we use AutoDock Vina extension in pymol and dock the protein with D-galactonate and DNA. It will give us several possible conformations(10 reiterations) of D-galactonate, with their base score written with them. We choose the one with the minimum base score. To double-check the structure we calculate the RMSD and RMSF score of all the structures and choose the with the best RMSD and RMSF score among all of them. After that all the required hydrogen atoms were added via the water box simulation, and to balance the charges a Mg atom is also added in the system by using the “trajconv” command of gromacs. So, the system now contains two proteins (dimer of DgoR), two nucleic acid (DNA), two ligands (D-galactonate) and a Mg atom.

Below is the figure of the docked modelled DgoR (Docked with D-galactonate and DNA). Different colours in the protein shows different residues present in DgoR.

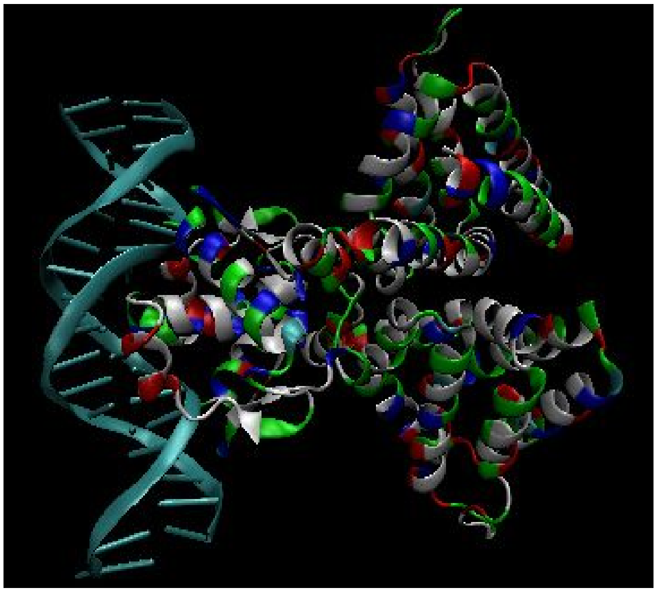


Figure 2.4(DgoR-DNA complex)

3. Result and analysis:

3.1 Structure:

We analysed the ligand-protein interaction. Below are residue name and Id of the residues interacting with ligand(D-galactonate) in both protein chains:-

Residue Id.		Residue name	Residue Id.		Residue name
4	→	ILE	6	→	THR
7,8	→	THR	19	→	ARG
19	→	GLU	21	→	GLU
24	→	SER	25	→	ALA
25	→	ALA	26	→	SER
28	→	ASN	29	→	ASN
30	→	LYS	34	→	LYS
39	→	TYR	38	→	TYR
43	→	ARG	44	→	ARG

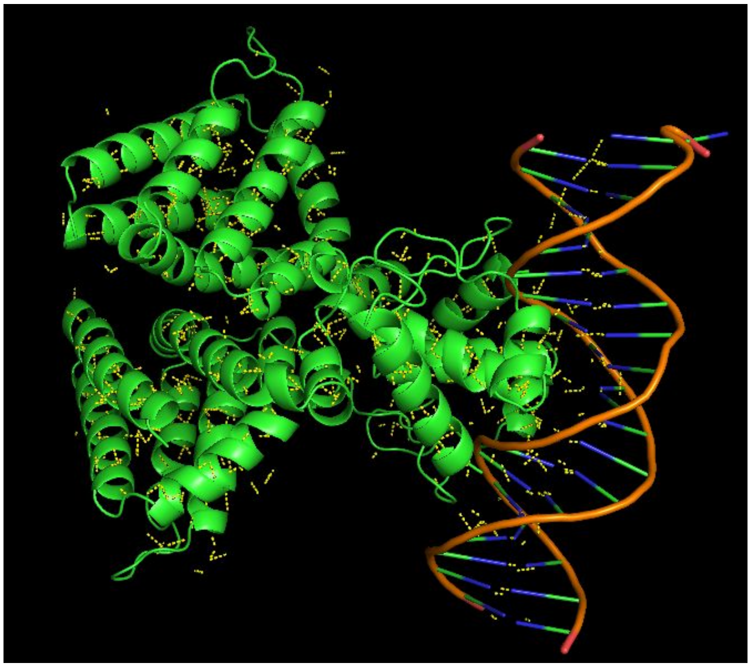


Figure 2.5 (polar contacts (shown as dotted yellow lines))

3.2 Simulating the protein over time:

The time-evolving coordinates of a system are called as trajectory of the protein. They are obtained in simulations of molecular systems. Trajectory files (with extensions such as .xtc and .dcd) are typically large binary files that contain the time varying atomic coordinates for the system (stored like a matrix). Each set of coordinates corresponds to one frame in time. Trajectory files do not contain structural information as found in structure files such PSF, PDB, GRO, TPR, etc.. So to analyze a molecular structure over time we must add the structure file along with the trajectory file. We generated a trajectory for DgoR of 50000 frames in XTC (most compatible and portable) and DCD format (by using CatDCD).

```
catdcd -o wt-dgoR-holo-dt10.dcd -otype dcd -xtc wt-dgoR-holo-dt10.xtc                (2.4.1)
```

PSF and CRD files were also created using ParmEd using this method, while converting the gromacs topology file (.top) to CHARMM topology file (.psf)

3.3 Root mean square fluctuation (RMSF):

It is the measure of the distance between the atoms over a period of time with respect to reference frame. We can also say that it is the calculation of individual residue movement during a simulation. RMSF per residue is typically plotted vs. residue number, which will tell us about the amino acids that contributes most to a molecular motion in a protein during a simulation over time (14).

$$\mathrm{RMSF}_i = \left[\frac{1}{T} \sum_{t_j=1}^{T} |\mathbf{r}_i(t_j) - \mathbf{r}_i^{\mathrm{ref}}|^2 \right]^{1/2} \qquad (3.1.1)$$

We calculated the RMSF for the C-alpha atoms of the protein and DNA.

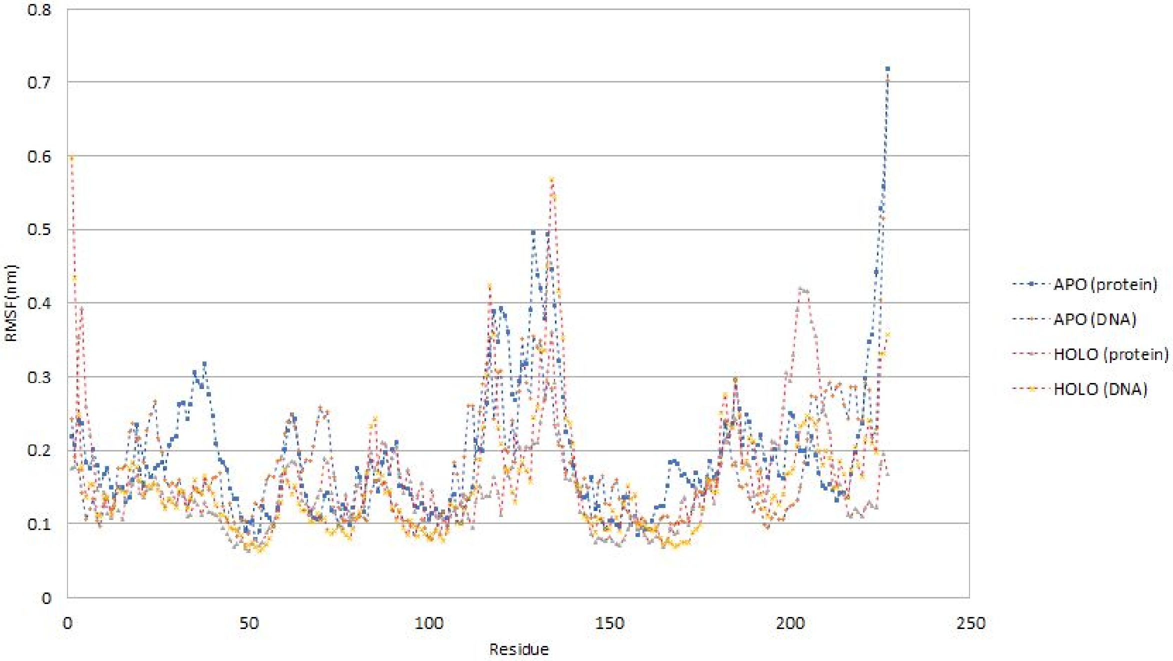


Figure 3.1 (RMSF)

RMSF values shows normal flactuaion pattern, other then that the residues between resid 120-130 shows a peak for C-alpha atoms of both DNA and protein. Average RMSF value increases slightly with respect to time.

Terminal residues usually shows higher RMSF values other then this we see that residues around Resid 120-130 show higher RMSF value (higher fluctuations) than other residues in both the forms. Other than these cases (in the case of the residues THR and LEU in starting) the residues in APO form show relatively higher fluctuations than those in HOLO form.

3.4 Root mean square deviation (RMSD):

For RMSD the average is taken over the particles, giving time specific values with respect to the reference frame. Root-mean-square deviation (RMSD) is the measure of the average distance between the atoms usually the backbone atoms of the protein. RMSD values representation mainly for analyzing stability of protein and predicting conformational changes of protein(14). Low RMSD of the binding pose with respect to the reference structure implies that particular binding pose is good.

$$\mathrm{RMSD}(t) = \left[\frac{1}{M}\sum_{i=1}^{N} m_i |\mathbf{r}_i(t) - \mathbf{r}_i^{\mathrm{ref}}|^2\right]^{1/2} \qquad (3.2.1)$$

where r_i is the distance between atom i and a reference structure or the mean position of the N equivalent atoms.

The most popular estimator of structural similarity is the root-mean-square distance (RMSD) between equivalent atoms, computed after optimal superposition of the two structures that are compared(14). It is known that RMSD values do not depend only on conformational differences but also on other features, for example the dimensions of the structures that are compared.

We have done RMSD on the backbone atoms of the protein to check its stability over time (50000 frames).

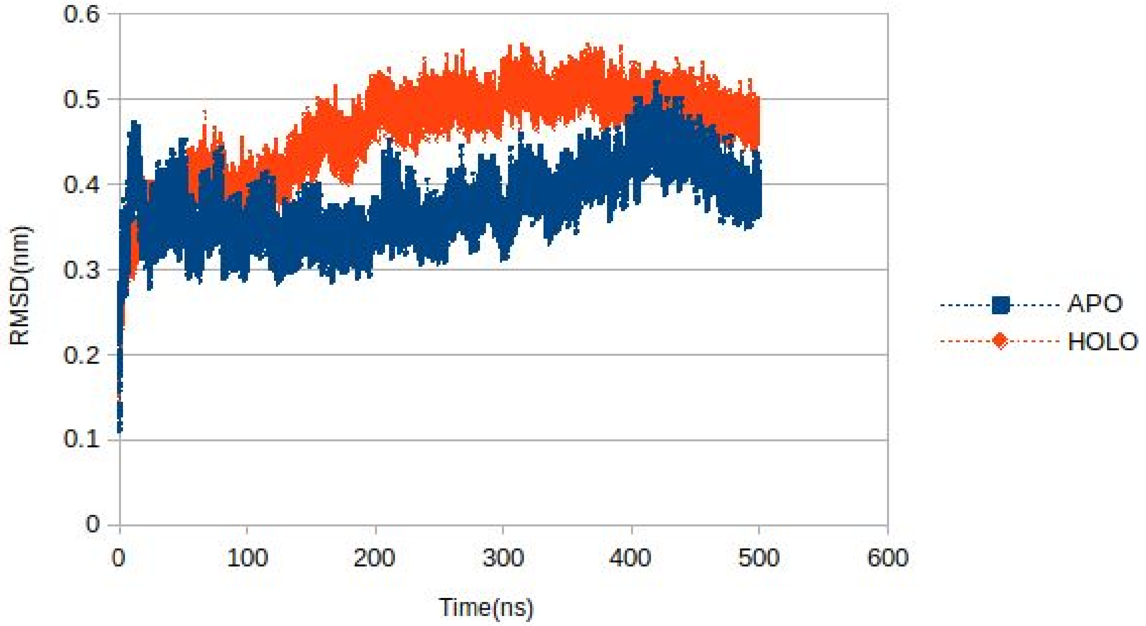


Figure 3.2 (RMSD)

RMSD value of the backbone atoms in both the form(APO and HOLO) increases with respect to time. The average RMSD value of backbone atoms over time in APO form is relatively less than those in HOLO form.

3.5 Native contact analysis:

Understanding the mechanism by which proteins fold to their native structure is a central problem in protein science. Realizing that the contacting residues which are far apart in the protein sequence but close together in the 3-D space are important for protein folding, contacts are widely categorized as short-range, medium-range, and long-range (15). Short-range contacts are those separated by 6–11 residues in the sequence; medium-range contacts are those separated by 12–23 residues, and long-range contacts are those separated by at least 24 residues.Interactions between residues that are in contact in the folded state (native interactions) play an important role in folding (16). The fraction of native contacts, is a natural choice of reaction coordinate for most of the models in which the energy is strongly correlated with *the native contacts*. We performed the native contact analysis on the short ranged contacts (within 5 angstrom) in MDAnalysis. MDAnalysis is an object oriented python toolkit distributed via the pip package. It is used to analyse the molecular dynamics trajectory created by CHARMM, GROMACS, AMBER, NAMD and any other software package using the similar force fields(21). It allows one to read molecular dynamics trajectories and access the atomic coordinates through NumPy arrays. This provides a flexible and relatively fast framework for complex analysis tasks(17).

We plotted the probability distribution function plot (Figure3.3) of native contact analysis using numpy and panda extension of python.

Conclusion:

In HOLO form (the one that contain D-galactonate as ligand) the most recurring fraction of native contacts is around 0.89 while is APO form the fraction of native contacts seems closer to 0.90 and a little farther from 0.89.

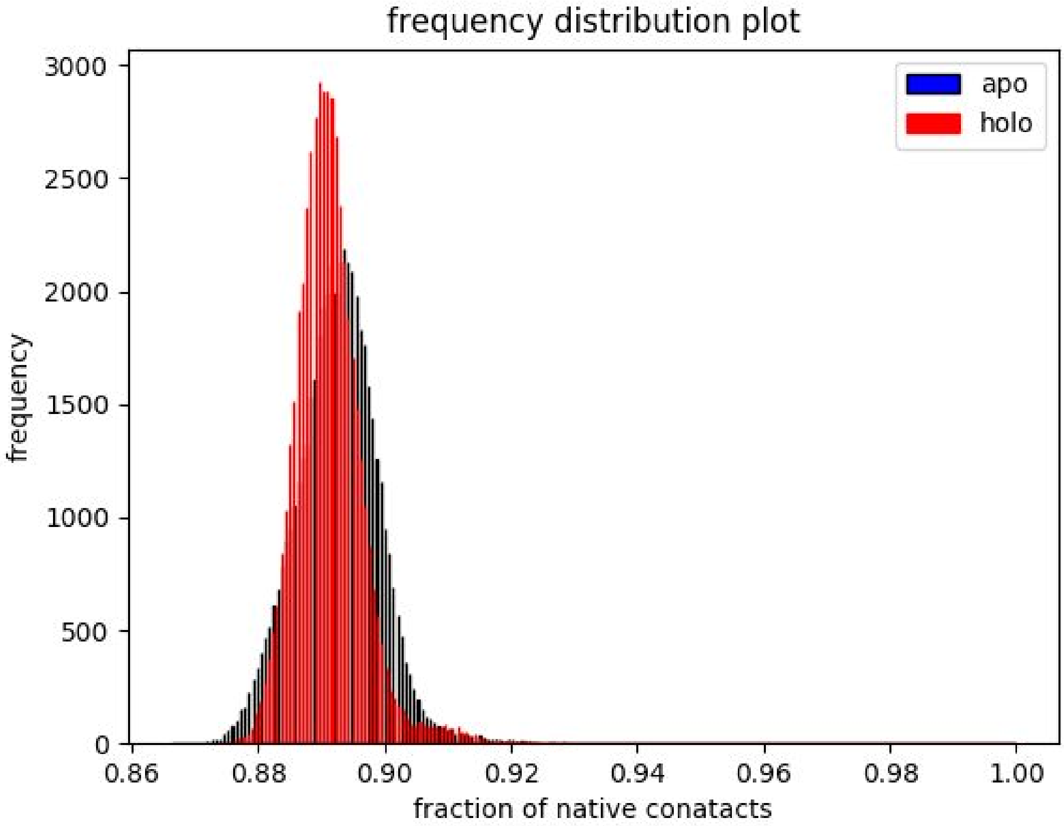


Figure 3.3 (Native contacts)

We used the following python code (in MDAnalysis) for calculating the native contacts within 5 angstrom:-

```
import numpy as np

import matplotlib.pyplot as plt

import MDAnalysis as mda

from MDAnalysis.analysis import contacts

from MDAnalysis.tests.datafiles import PSF,DCD
```

```
u = mda.Universe('charmm_holo.psf','holo.dcd')

q1q2 = contacts.q1q2(u, 'name CA', radius=6)

q1q2.run()

average_contacts = np.mean(q1q2.timeseries[:, 1])

print('average contacts = {}'.format(average_contacts))

fig, ax = plt.subplots()

ax.plot(q1q2.timeseries[:, 0], q1q2.timeseries[:, 1])

ax.set(xlabel='frame', ylabel='fraction of native contacts',

                             title='Native    Contacts,    average    =
{:.2f}'.format(average_contacts))

fig.show()

#taking output into a text/csv file

filename = 'nca_fre.csv'

out_fl_name = "/home/kyadav/Documents/final/DGOR/" + filename

out_fl = open(out_fl_name, 'w+')

out_header = "frames, nca"

out_fl.write(out_header)

for i in range(len(q1q2.timeseries[:, 0])):

    out_fl.write('\n')

             out_fl.write(str(q1q2.timeseries[:,    0][i])    +    ","    +
str(q1q2.timeseries[:, 1][i]))

out_fl.close()
```

(3.3.1)

3.6 Methodology to find the non-polar contacts in system:

We use the following program to find the ligand-protein and DNA-protein interaction in the HOLO system:

```
proc residueContactPairs { cutoff sel1 sel2 } {
  set cl [ measure contacts $cutoff $sel1 $sel2 ]
  if { $cl == {} } {return {} }
  set l1 [lindex $cl 0]
  set l2 [lindex $cl 1]
  set n [llength $l1]
  array set pairs {}
  for {set i 0} {$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]
    $tmp delete
    if {$r1 !=$r2} {
           if { ! [info exists pairs($r2,$r1)] } {
        set pairs($r1,$r2) 1 }}}
       set plist {}
  foreach p [array names pairs] {
    set pl [split $p ,]
    lappend plist $pl }
    return $plist }
mol new wt-dgoR-holo-dt10.xtc type xtc waitfor all
mol addfile wt-dgoR-holo-reference.gro type gro molid 0
set outfile [open "GAL-DgoR.dat" w]
set num_steps [molinfo top get numframes]
```

```
for {set frame 0} {$frame < $num_steps} {incr frame} {
 set sel1 [atomselect top "resname GAL 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
```

3.7 DNA-Protein interaction:

DNA-binding proteins are proteins that have DNA-binding domains and thus an affinity for single or double stranded DNA. Protein-DNA interactions are essential components of all biological systems, fundamental to almost all biological processes(18). Specificity in DNA-protein interactions comes from protein recognition of the linear order of base pairs through hydrogen bond and salt bridge contacts through the major and minor grooves.

DgoR with D-galactonate (HOLO form) in its system shows less percent of close contacts or interaction of DNA with protein than the system with DgoR alone (i.e. APO form). For example as we seen in the figure below (Figure 3.4 and 3.5) 7DG in HOLO is not showing as much contact as 7DG in APO form.

In the below figure, comparatively lighter color represents those residues of the protein that came in more contact of the DNA, while the darker colour represents the one with less contacts.

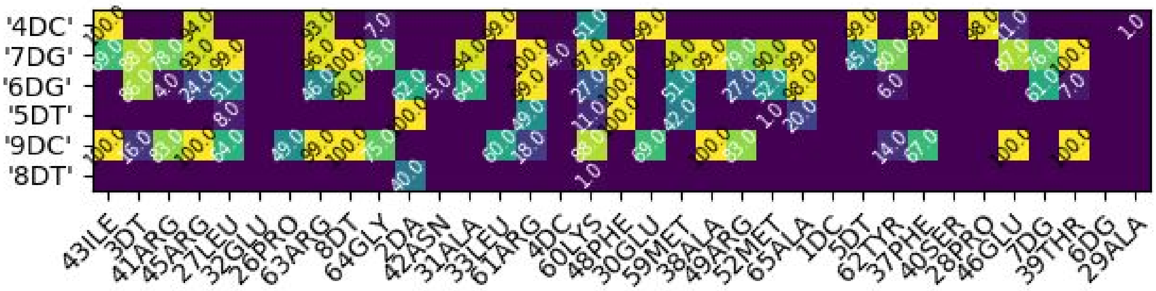


Figure 3.4 (DNA-protein interaction in APO form)

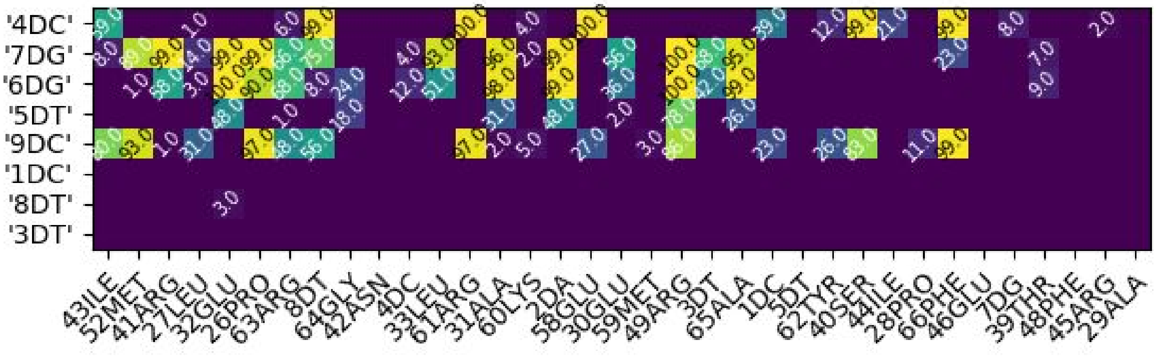


Figure 3.5 (DNA-protein interaction in HOLO form)

3.8 Ligand protein interaction:

A ligand is a comparatively small molecule that is able to bind itself to proteins by weak bond interactions such as ionic bonds, hydrogen bonds, Van der Waals interactions, etc... In some particular cases, a ligand will also serves as a signal triggering molecule. The docking is usually a reversible reaction (dissociation). Some of the things that we need to consider while observing Ligand-Protein interaction are Binding site (This site usually exhibits specificity to ligands, the weak interactions within the primitive structure of the protein such as the side chain initiates the response that results in the ligand going for a specific binding site), Induced fit (it is a concept based on which we assume that the ligand is mobile/flexible and not rigid in its movements) and dissociation constant(it can be defined as the tendency for the ligand to the binding site, measured by the concentration of ligand over the concentration of the ligand-protein complex).

We are going to check the ligand-protein interaction of the complex for the residues of the protein that are within 5 angstrom of distance to the ligand (D-galactonate).

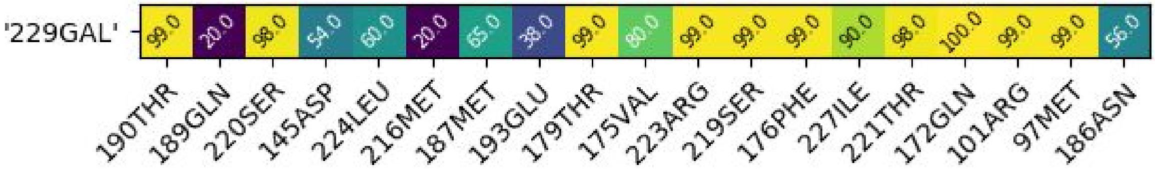


Figure 3.6 (Ligand-protein interactions)

Here we see the residues of the protein that are interacting with D-galactonate. In the above figure, comparatively lighter color represents those residues of the protein that came in more contact of the ligand, while the darker colour represents the one with less contacts. These are the residues that came in contact with D-galactonate over time: Threonine(THR), Glutamine(GLN), Serine(SER), Aspartic Acid(ASP), Leucine(LEU), Methionine(MET), Glutamic acid(GLU), Valine(VAL), Arginine(ARG), Phenylalanine(PHE), Isoleucine(ILE), Asparagine(ASN).

Bibliography:

1. Determinants of enhancer and promoter activities of regulatory elements, Robin Andersson and Albin Sandel
2. Phillips, T. (2008) Regulation of transcription and gene expression in eukaryotes. Nature Education 1(1):199
3. Molecular and Functional Insights into the Regulation of D-Galactonate Metabolism by the Transcriptional Regulator DgoR in Escherichia coli. Bhupinder Singh,a Garima Arya,a Neeladrita Kundu,a Akshay Sangwan,a Shachikanta Nongthombam,a Rachna Chabaa
4. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. 2019 Jan 8; 47(D1):D1102-1109. doi:10.1093/nar/gky1033. [PubMed PMID: 30371825]
5. Differences Between Apo and Three Holo Forms of the Intestinal Fatty Acid Binding Protein Seen by Molecular Dynamics Computer Calculations, Thomas B.Woolf, Alan Grossfield, Michael Tychko, Volume 78, Issue 2, February 2000, Pages 608-625.
6. Rancour NJ, Hawkins ED, Wells WW. 1979. Galactose oxidation in liver. Arch Biochem Biophys 193:232-41.
7. Molecular modeling and simulations, I.Nezbeda*, J.Jirsák*, F.Moučka*, 2017
8. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers, Mark James Abrahama, TeemuMurtolad, RolandSchulz, SzilárdPáll, Jeremy C.SmithbcBerkHess, ErikLindahlad
9. CHARMM: The Biomolecular Simulation Program, B.R. Brooks,1,* C.L. Brooks, III,2,* A.D. MacKerell, Jr.,3,* L. Nilsson,4,* R.J. Petrella,5,6,* B. Roux,7,* Y. Won,8,* G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A.R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R.W. Pastor, C.B. Post, J.Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D.M. York, and M. Karplus5,9,*
10. D.A. Case, R.M. Betz, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao and P.A. Kollman (2016), AMBER 2016, University of California, San Francisco

11. https://parmed.github.io/ParmEd/html/_modules/parmed/structure.html, ParmEd

12. The FadR.DNA complex. Transcriptional control of fatty acid metabolism in Escherichia coli, Xu, Y., Heath, R.J., Li, Z., Rock, C.O., White, S.W. (2001) J.Biol.Chem. 276: 17373-17379

13. www.rcsb.org Protein Data bank.

14. Rmsd/Rmsf Analysis | Biochemcore 2018

    https://ctlee.github.io/BioChemCoRe-2018/rmsd-rmsf/

15. Native contacts determine protein folding mechanisms in atomistic simulations Robert B. Best, Gerhard Hummer, and William A. Eaton

16. Protein Residue Contacts and Prediction Methods, Badri Adhikari and Jianlin Cheng

17. R. J. Gowers, M. Linke, J. Barnoud, T. J. E. Reddy, M. N. Melo, S. L. Seyler, D. L. Dotson, J. Domanski, S. Buchoux, I. M. Kenney, and O. Beckstein. MDAnalysis: A Python package for the rapid analysis of molecular dynamics simulations. In S. Benthall and S. Rostrup, editors, Proceedings of the 15th Python in Science Conference, pages 98-105, Austin, TX, 2016. SciPy, doi:10.25080/majora-629e541a-00e.

18. B. Webb, A. Sali. Comparative Protein Structure Modeling Using Modeller. Current Protocols in Bioinformatics 54, John Wiley & Sons, Inc., 5.6.1-5.6.37, 2016.

19. Molecular Docking: A powerful approach for structure-based drug discovery Xuan-Yu Meng,1,2 Hong-Xing Zhang,1,* Mihaly Mezei,3 and Meng Cui2,*Curr Comput Aided Drug Des. 2011 Jun 1; 7(2): Pages 146–157.

20. Rational Structure-Based Drug Design,

    Varun Khanna, Shoba Ranganathan, Nikolai Petrovsky, Sep 2018.