UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Posgrados

Interaction of ZIKV NS5 and STAT2 examined by Molecular Modeling, Docking and Simulations Studies

Ángel Gerardo Armijos Capa

Miguel Ángel Méndez, PhD.
Director de Trabajo de Titulación

Trabajo de titulación de posgrado presentado como requisito
para la obtención del título del Máster en Química

Quito, 5 de diciembre de 2018

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Posgrados

HOJA DE APROBACIÓN DE TRABAJO DE TITULACIÓN

Interaction of ZIKV NS5 and STAT2 examined by Molecular Modeling, Docking and Simulations Studies

Miguel Ángel Méndez, PhD.
Director del Trabajo de Titulación

F. Javier Torres, PhD.
Director del Programa de Maestría en Química

César Zambrano, PhD.
Decano del Colegio Politécnico

Hugo Burgos, PhD.
Decano del Colegio de Posgrados

Quito, 5 de diciembre de 2018

©Derechos de Autor

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior.

Firma del estudiante: ______________________________

Nombre: Ángel Gerardo Armijos Capa

Código de estudiante: 00140980

C. I.: 1722073341

Lugar, Fecha: 5 de diciembre de 2018

To my parents and brothers

Acknowledgment

To my advisor Miguel Ángel Méndez and all QCT's group for their support to perform this research.

Resumen

El virus Zika (ZIKV) codifica para la proteína NS5, la cual se conoce como un antagonista potente y específico de la señalización de interferón (IFN). ZIKV NS5 se ha asociado con la degradación proteosomal del transductor de señal y el activador de la transcripción 2 (STAT2), aunque el mecanismo completo aún se desconoce ya que los estudios experimentales sugieren que dominios diferentes a Mtase contribuyen a la degradación de STAT2. Se ha empleado estudios de modelado molecular, acoplamiento y dinámica molecular para explorar en detalle las características dinámicas y estructurales del complejo NS5, STAT2 y NS5-STAT2, así como los residuos clave involucrados en la interacción NS5-STAT2. En este estudio, se ha validado un modelo tridimensional de STAT2 (C-score = -0.62) que no se ha informado experimentalmente. Del mismo modo, un complejo NS5-STAT2 se ha detallado entre varios modelos a través de la energía total de estabilización de -77.942 kcal·mol^{-1} y una energía libre de enlace de Gibbs de -4.30 kcal·mol^{-1}. Los resultados han revelado que la interacción se limita a tres dominios conocidos como N-terminal de STAT2 y Mtase-Thumb de NS5 que se ubican en las regiones ordenadas de ambas proteínas. Los residuos clave que intervienen en la superficie de interacción con la frecuencia más alta se estabilizan mediante interacciones electrostáticas, interacciones hidrófobas, puentes salinos e interacciones iónicas. Por lo tanto, nuestro hallazgo respalda las observaciones preliminares experimentales informadas en la literatura y ayudará en los esfuerzos de diseño de medicamentos contra ZIKV NS5.

Palabras clave: Zika virus · NS5 · STAT2 · Dinámica Molecular · Docking.

Abstract

Zika virus (ZIKV) encodes NS5 protein which is known as a potent and specific antagonist of Interferon (IFN) signaling. ZIKV NS5 has been associated with the proteosomal degradation of signal transducer and activator of transcription 2 (STAT2), although the complete mechanism is still unknown since experimental studies suggest that additional regions to Mtase domain may contribute the STAT2 degradation. It has been employed molecular modeling, docking and MD studies to explore into detail the structural and dynamic features of NS5, STAT2 and NS5-STAT2 complex as well as the key residues involved in NS5-STAT2 interaction. In this study, it has been validated a three-dimensional model of STAT2 (C-score=-0.62) which has not been reported experimentally. Likewise, a docked complex NS5-STAT2 has been detailed among several models through the total stabilizing energy of -77.942 kcal·mol^{-1} and a Gibbs binding free energy of -4.30 kcal·mol^{-1}. The outcomes have revealed that interaction is limited to three domains known as N-terminal from STAT2 and Mtase-Thumb from NS5 locating in the ordered regions of both proteins. Key residues involved in the interaction surface with the highest frequency are stabilized by electrostatic interactions, hydrophobic interactions, salt bridges and ionic interactions. Therefore, our finding supports the experimental preliminaries observations reported in the literature and will help in the drug design efforts against ZIKV NS5.

Keywords: Zika virus · NS5 · STAT2 · Molecular Dynamics · Docking.

Contents

Acknowledgment **5**

Resumen **6**

Abstract **7**

1 Introduction **16**

2 Literature review **20**

2.1 Protein structure modeling . . . 20

2.1.1 Modeling method . . . 20

2.1.2 Step 1: Searching structures . . . 21

2.1.3 Step 2: Selecting templates . . . 21

2.1.4 Step 3: Building model . . . 22

Templates-dependent modeling . . . 22

Templates-independent modeling . . . 22

Step 4: Refinement . . . 23

Step 5: Evaluation . . . 23

2.2 Molecular dynamics simulation on biomolecules . . . 23

2.2.1 Potential energy function . . . 25

2.2.2 AMBER force field . . . 27

2.2.3 Algorithms for MD simulations . . . 28

2.2.4 Periodic boundary conditions . . . 29

2.2.5 Ensembles in MD . . . 30

2.2.6 MD at constant temperature . . . 30

Berendsen thermostat . . . 31

Nosé-Hoover thermostat . . . 31

2.2.7 MD at constant pressure . . . 32

Berendsen barostat . . . 32

Parrinello-Rahman barostat . . . 33

2.2.8 Simulation methods . . . 33

2.3 Docking . . . 35

2.3.1 Docking procedures . . . 36

Searching of first docked complexes . . . 36

Rescoring . . . 37

Flexibility on the structure . . . 37

2.3.2 Algorithms use in docking . . . 37

Systematic search algorithms . . . 37

Stochastic search algorithms . . . 38

Deterministic search algorithms . . . 38

2.3.3 Protein-protein docking . . . 39

2.4 Gibbs binding free energy . . . 40

2.4.1 Methods of estimation for Gibbs binding free energy . . . 41

Thermodynamic Integration (TI) . . . 41

Free Energy Perturbation (FEP) . . . 41

Molecular mechanics force field . . . 42

2.4.2 Time-consuming process of ΔG_{Bind} . . . 43

First-principles methods . . . 43

Semiempirical methods . . . 43

Empirical methods . . . 44

3 Methods **45**

3.1 Sequence analysis of NS5 and STAT2 45

3.2 Molecular modeling of NS5 and STAT2 46

3.3 Structures quality check and structural alignment 46

3.4 Molecular dynamics simulation of NS5 and STAT2 47

3.5 NS5 and STAT2 ensembles for protein-protein docking 48

3.6 Protein-protein interaction and Binding energy of binary complex NS5-STAT2 49

3.7 Electrostatic Calculations 51

4 Results and discussion **53**

4.1 Analysis of amino acid sequences 53

4.2 Molecular modeling and validation of STAT2 57

4.3 MD simulation of NS5 and STAT2 61

4.4 Analysis of docking complex NS5-STAT2 65

4.5 Contact map and electrostatic analysis 71

4.6 MD simulation of NS5-STAT2 complexes 73

4.7 Interaction NS5-STAT2 78

5 Conclusions **81**

Bibliography **83**

List of Tables

4.1 Prediction of secondary structure and solvent access to the amino acid sequence of NS5 and STAT2 through RaptorX property prediction and GOR4. 56

4.2 Analysis of global quality of structure, secondary structure and free energy of folding to the initial and final coordinates with 10 ns of simulation of NS5 and STAT2 through Verify-3D, ERRAT, and VADAR. 59

4.3 Results of total stabilizing energy and normalized energy per residue for the ten models of ClusPro and Pydock computed by FoldX. 68

4.4 Results of total stabilizing energy and normalized energy per residue for ten ClusPro's models based on the functional domains of NS5 and STAT2. The functional domains involved in the interaction of complex were the N-terminal from STAT2 and Mtase and Thumb from NS5. 70

List of Figures

2.1 Basic algorithm of MD simulation. The global steps such as energy calculation, estimation of forces, numerical integration and trajectory. E_{pot} is potential energy, t is simulation time, dt is iteration time, F is force component, a is acceleration, m is atom mass, and v is velocity. 25
2.2 Procedure of docking. 36

4.1 NS5 and STAT2 amino acid composition based on their sequence. 54
4.2 Analysis of disordered regions and binding residues in target sequence proteins through DISOPRED. (a) NS5 shows disordered regions in 1-6, 269-278, 460-466 and (b) STAT2 in 130-141, 185, 326, 371-373, 375-394, 410-423, and 706-851 of the sequence. Peaks (blue) over 0.5 confident score are considered regions with high level of disordered. Peaks (orange) are regions with binding residues. 55
4.3 Ramachandran plot of STAT2 model by ITASSER and Phyre2. (a) Phyre2's model (b) I-TASSER's model (c) STAT2's model with 10 ns MD simulation. The most favored regions are red. Allowed, generously allowed, and disallowed regions are yellow, pale yellow, and white, respectively. High stereo-chemical quality is considered when any residue lies in the disallowed region. 59

4.4 Ramachandran plot of NS5 model (PDB code: 5TMH). (a) NS5 without MD simulation (b) NS5 with 10 ns MD simulation. The most favored regions are red. Allowed, generously allowed, and disallowed regions are yellow, pale yellow, and white, respectively. High stereo-chemical quality is considered when any residue lies in the disallowed region. 60
4.5 Three-dimensional structures of (a) NS5 and (b) STAT2 model. The functional domains of each model are shown in different color. NS5: Mtase = Methyltransferase domain (1-264 aa), Linker domain (265-275 aa), Extension (275-304 aa), Fingers domain (305-477 aa), Palm domain (478-714 aa) and Thumb domain (715-903 aa) STAT2: N-terminal domain (1-138 aa) involved in dimerization/-tetramerization, Coiled-coil domain (139-315 aa) involved in interaction with other proteins, DNA binding domain (316-485 aa), Linker domain (486-574 aa), SH2 domain (575-679 aa), Phosphotyrosyl tail segment (680-697 aa), and Transactivation domain (698-851 aa). 60
4.6 Structural comparison among the experimental fragments of STAT2 (PDB codes: 5OEN and 2KA4) with STAT2's model. (a) Fragment 2KA4 and STAT2's model (b) Fragment 5OEN and STAT2's model. 61
4.7 RMSD for NS5 and STAT2 as a function of time. (a) RMSD evolution of NS5 (b) RMSD evolution of STAT2. 63
4.8 RMSF of the backbone Cα atoms versus the number of residues present in each sequence both (a) NS5 as (b) STAT2. 64
4.9 Radius of gyration (Rg) and solvent Accessible Surface (SAS) for NS5 and STAT2 as function of time. (a) Rg evolution of NS5 (b) Rg evolution of STAT2 (c) SAS evolution of NS5 (d) SAS evolution of STAT2 during 50 ns. 64
4.10 The first ten models of NS5-STAT2 complex obtained by (a) ClusPro and (b) Pydock. 69
4.11 Results of PPCheck web server for total stabilizing energy in ClusPro's models. 69

4.12 Plot of total stabilizing energy from the query domains (marked with red) with the known interactions (marked with blue) from other complex obtained from PIMA. 70

4.13 Results obtained through PISA web server. (a) Interface area (Å^2) and (b) solvation free energy achieved by the formation of the interface for ClusPro and Pydock. 70

4.14 Results of Gibbs free energies of binding for ten ClusPro's models calculated by FoldX software. 71

4.15 Contact map of NS5 and STAT2 shows the intermolecular contacts at reducing distances which are red=7Å, yellow=10Å, green=13Åand blue=16Å. The circles identify the domains related such as N-terminal/Mtase, N-terminal/Thumb and Thumb/DNA binding. 72

4.16 A view of electrostatic surface potential for (a) NS5-STAT2 complex, (b) STAT2 and (c) NS5. Red is negative charge, blue is positive charge, and white is neutral. 73

4.17 RMSD of bound forms of NS5 and STAT2 compared with the unbound form of each protein. (a) RMSD of NS5 (b) RMSD of STAT2 as unbound and bound forms 75

4.18 RMSF backbone Cα atoms of bound forms of NS5 and STAT2 compared with the unbound form of each protein. (a) RMSF of NS5 (b) RMSF of STAT2 as unbound and bound forms 75

4.19 RG of bound form of NS5 and STAT2 compared with the unbound form of each protein. (a) RG of NS5 (b) RG of STAT2 as unbound and bound forms 76

4.20 SAS of bound form of NS5 and STAT2 compared with the unbound form of each protein. (a) SAS of NS5 (b) SAS of STAT2 as unbound and bound forms 76

4.21 Frequency of residues in the docked complex that have interactions in the interface. High frequency of residues in (a) NS5 and (b) STAT2 that showed any type of interaction such as electrostatic,hydrophobic, salt bridges or ionic interaction. 80

Chapter 1

Introduction

Since 2015, the outbreak of Zika virus (ZIKV) in Central and South America has become a worldwide health concern. Therefore, the World Health Organization (WHO) has declared ZIKV infection as a disease of international public health emergence because during pregnancy has been associated with neurological disorders. It has been associated with congenital diseases that may cause severe disabilities such as microcephaly, Guillain-Barré syndrome, and destruction of neural cells [1], [2], [3], [4], [5].

Zika virus is a mosquito-borne flavivirus that has a genome of single-strand positive RNA of 10 kb that encodes ten proteins classified into three namely the envelope (E), membrane precursor (PrM), and capsid (C) which contribute to the viral particles, and seven nonstructural (NS) proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5) which contribute to viral replication [1], [2], [4], [5]. NS1, NS3, and NS5 are large and highly-conserved proteins while NS2A, NS2B, NS4A, and NS4B are small and hydrophobic [4], [6]. Among NS proteins, NS5 is the largest viral protein with a weight of ∼103 kDa and consists in two principal domains known as the methytransferase (Mtase) domain containing the first 264 amino acid at the N-terminal portion of NS5, followed by a short linker that connects to the RNA-dependent RNA polymerase (RdRp) domain (275-903 aa) [5], [7]. The Mtase domain is related with the decrease of host innate immune response and promotes the translation of the polyprotein through the addition of 5' RNA cap structure. The RdRp domain is required for the initiation and elongation of RNA synthesis [2], [7]. The viral response in the human cells is set up by intracellular pattern recognition receptors (PRRs) which are proteins that recognize viral pathogen as viral RNA, DNA or protein in the type I interferon (IFN) signaling. This recognition induces activation of innate immune signaling, leading to the up-regulation of IFN-stimulated genes (ISGs). Various strategies of IFN antagonism had been described for flaviviruses such as Dengue virus (DENV), West Nile virus (WNY), Yellow Fever virus (YFV) [8]. However, flaviviruses share replication strategies based on the formation of a polyprotein which can inhibit transcriptional activation of IFNs and ISGs during virus infection [9], [10], [11]. In this vein, NS5 is considered as a potent and specific antagonist of type I IFN signaling [4], [8], [12].

Signal Transducer and Activator of Transcription 2 (STAT2) is a protein within the pathway of type I IFN in mammalian cells that has the capability to transduce signals from the cell membrane to the nucleus to active gene transcription as ISGs [13]. STAT2 is stimulated by cytokines, including interferons and interleukins as IFNα and Janus Kinase (JAK) proteins which through the phosphorylation of tyrosine (Tyr) residue allows access to STAT2 [14]. A phosphorylated complex is formed with STAT2 which is bound to interferon receptor 9 (IRF 9) and STAT1, and forms a ternary complex known as IFN-gene factor 3 (ISGF3). ISGF3 complex translocates into the nucleus which binds a section of DNA sequence called IFN-stimulated response element (ISRE) [14]. Finally, this latter permits the transcription of ISGs genes which will block viral infection through antiviral proteins known as type I IFN [15]. In contrast, STAT2 can be inactivated by negative regulators as cytokines, phosphatases or other proteins in degradation pathway as for instance through ubiquitin-proteosome [13], [16]. STAT2 consists of an N-terminal domain (1-138 aa), Coiled-coil domain (139-315 aa), DNA binding domain (316-485 aa), Linker domain (486-574 aa), SH2 domain (575-679 aa), Phosphotyrosyl tail segment (680-697 aa) and Transactivation domain (698-851 aa). Moreover, STAT2 conserves a Tyr residue at the C-terminus that supplies the phosphorylation for its activation and generates intramolecular interaction with its dimer partner at the SH2 domain [13], [16]. However, it has also been reported the lacks of a second conserved phospho-amino acid residue at the

Transactivation domain, as it has been observed in others STAT protein [13].

ZIKV NS5 has been associated with inhibition of type I IFN during the host antiviral response, because it may promote the proteosomal degradation of STAT2. This latter brief description has been substantiated by different experimental studies which have determined that strains of ZIKV antagonize type I IFN where NS5 reduces the STAT2 level and prevents the translocation of STAT2 from cytoplasm to nuclei in immunoprecipitation essays in 293T cells [8]. Consequently, there is inhibition of the IFN induction of the ISG [8]. In others test, STAT2 levels have been compared in cells express each of the two functional domains of NS5. The comparison has showed that with only the Mtase domain expressed, the levels of STAT2 degradation were higher, while with only RdRp domain expressed, the levels of STAT2 degradation were negligible [11]. On the other hand, the Mtase domain in full-length NS5 has showed incapability to induce the STAT2 degradation, suggesting that other regions of the protein NS5 may contribute to degradation mechanism [11]. Therefore, a deficiency of STAT2 carries that patients will have a considerable susceptibility to viral infections since there is a deregulation in response genes of IFN pathway with antiviral activity [14]. Another essays where have been reviewed the interaction with other possible proteins in the pathway of type I IFN have shown that reduced STAT2 levels associated to ZIKV infection, are independently of the presence of a protein such as ubiquitin ligase UBR4 or ubiquitin-specific protease USP 18, hence the interaction between STAT2-NS5 is directly related with degradation of STAT2, although the complete mechanisms are still unknown [8], [9], [10], [11].

Additionally, a factor that must be taken into account in the interaction among proteins is the phosphorylation. In NS5 may play an important role in the association with cellular kinases [17]. The residues such as serine (Ser), threonine (Thr) or tyrosine (Tyr) are phosphorylated to convert NS5 from a non-phosphorylated to phosphorylated state [18], [19]. In fact, experimental studies performed in nuclei of infected cells suggest that NS5 phosphorylation may regulate the viral replication, nuclear transport, the expression of host genes in response to viral infection or may control the interaction with proteins such as NS3 [17], [20]. Indeed, it has been demonstrated that a hyperphosphorylated state of NS5 at Ser affects the interaction with NS3 [19]. Otherwise, in order to form homodimers or heterodimers, STAT2 needs to be activate by phosphorylation on Tyr [9], [21]. Studies have shown that phosphorylated regions of amino acid 148-324 in STAT2 and 217-377 in IRF9 are employed to interact and form a complex [22]. Furthermore, the binding between DNA and STAT2 takes place after Tyr phosphorylation [23]. Therefore, the phosphorylation processes determine the function in the proteins, however in this study will be used non-phophorylated forms of NS5 and STAT2 because the computational cost is low.

A complete understanding of the mechanism of interaction between NS5-STAT2 complex has proved to be difficult to obtain experimentally. However, the combination of different computational tools such as homology modeling to predict validate structures, molecular docking which is used to predict the binding interactions between ligand and receptor, and molecular dynamic (MD) simulation that has been used to describe the stability, strength and flexibility of binding between complex permit to develop a complex structure with a clear binding mechanism that could lead to significant advance in the understanding of NS5-STAT2 complex and would open a gate to future studies in the potential therapeutic intervention to treat ZIKV infections [24], [25]. Besides, the key binding site locations between NS5 and STAT2, and whether they participate or not in the formation of the complex may be elucidated by using a computational approach. The aim of the present study is to elucidate the potential role and mechanism of the NS5-STAT2 interaction involved in the process of infection by ZIKV pathway of type I IFN. For this purpose, several computational methods such as homology modeling, docking studies and Molecular Dynamics (MD) simulation were used.

Chapter 2

Literature review

2.1 Protein structure modeling

Three-dimensional structure of proteins describes countless features over their biological and physical-chemistry functions. Great efforts have been made to determine the protein structure using experimental methods. However, these approaches are not always applicable or the cost and time consuming are still elevated. Therefore, computational methods has been developed to predict three-dimensional structures from its amino acid sequence therewith has someway allowed to overcome the divergence between the number of sequences and three-dimensional models [26], [27].

2.1.1 Modeling method

Four different strategies have been used to produce a three-dimensional model. The first is homology modeling which uses experimental structures of related protein as template in order to model the target protein. It is usually most accurate approach. Second is the fold recognition and threading methods which are employed when a target sequence is totally new with respect to proteins with known structure. Third is based on *ab initio* methods. The models are generated using physics properties or through only information of known structures. Fourth is a group of methods that combine a set of computational and experimental information and use the principles of the three approaches before mentioned

[27]. Several programs and web servers have implemented this latter method through five sequential steps using as base the comparative modeling method.

2.1.2 Step 1: Searching structures

Searching of structures that are related with a target sequence are performed in different ways. Pairwise sequence-sequence comparison and fold assignment is implemented in programs as BLAST that compares the target sequence with each sequence in the databases, but it is not efficiency. In order to improve the sensitivity of searching, it is used the multiple alignment between the target sequence and multiple sequences. The program PSI-BLAST is usually employed to perform multiple alignment with a target sequences, for this purpose, it has implemented a heuristic search algorithm that looks for short motifs [26]. Likewise, the location of universal conserved motif among sequences has also been implemented with the profile-based Hidden Markov Models (HMM) algorithm. Other programs as COACH and FFAS03 increase the sensitivity of these approaches [26]. On the other hand, a second class of methods known as threading or fold assignment have been promoted to compare a sequence with a protein structure by a pairwise comparison. Here, it is used a library of 3D folds against the target sequences allowing to locate sequences that are barely related with known sequences [26], [27].

2.1.3 Step 2: Selecting templates

The step 1 generates a list of templates, but they need to be classified for the target sequence. One straightforward rule is the selection of structures with the highest sequence similarity. Multiple alignment or phylogenetic tree aids in the selection of templates since they increase the model accuracy. Likewise, complex methods of selection have been detailed using energy or scoring function which are much more accuracy. However, factors in the selection of templates such as quaternary interactions, ligands, pH or solvent should be considered when it is chosen a template [26].

2.1.4 Step 3: Building model

With the selection of a template, an initial model must be built, at this stage it is necessary to classify into templates-dependent and template-independent modeling [26].

Templates-dependent modeling

Different methods are proposed for modeling structures for instance assembly of rigid bodies, coordinate reconstruction, spatial restraints or combining structures. However, all of them are based on a template those features in the structure is employed as a framework to new model. For instance, an assembly of rigid bodies uses small number of rigid bodies from template to build a comparative model. The use of more template structures increases the accuracy of model. Another example is through spatial restraints in that the new model is built according to experimental-derived restrains. Alongside stereo-chemical restrains which are derived on bond lengths, bond and dihedral angles, and nonbonded contacts from molecular mechanics force field [26].

Templates-independent modeling

Target sequences that structurally have different regions or there is not information about a segment as surface loops with respect to template structures. These loops are considered functional sites as regions of binding for proteins and provided information about their local fold [26]. Besides, they usually represent gaps in the template when it is aligned with target sequence. Hence, it is appropriate to build the model without the aid of any template [26]. The modeling of loops are led in two approaches, which are fragment-based and *ab initio* modeling. The first uses a database of loops conformations where fragments are selected from a library. Fragments selected should satisfied geometrical restrains to fit inside of the target structure. However, this approach is limited, consequently searching methods are used to loop prediction such as molecular dynamics simulation, Monte Carlo simulation or self-consistent field optimization among others [26].

Step 4: Refinement

An initial model must be refined in order to improve the bond geometry or remove steric clashes. For that, it is used an energy minimization step by way of several approaches such as molecular mechanics force fields, molecular dynamics, Monte Carlo or genetic algorithms [27]. An example of refinement is using Monte Carlo sampling that is focused on regions with errors while remain structure (backbone and side chains) has been relaxed in an all-atom force field [27].

Step 5: Evaluation

This last step checks the possible errors that may have a new model. Hence, the quality of model can be examined from the sequence similarity with the template or with internal and external evaluations [26], [27]. Internal evaluation examines the model according to restrains used to build it as well as restrains from template and statistical observations. While external evaluation is only related with the use of scoring functions wherein parameters as energy is employed to classify models [26], [27].

2.2 Molecular dynamics simulation on biomolecules

Macromolecular structures have a key role in the biological functions which are based on the interactions and dynamic [28]. Hence, the functions of proteins are linked with conformational changes that take place on short time lapses [29]. The MD simulation of biomolecules (proteins or peptides) can be used to response a specific concern about individual particle motion or properties of a system as function of time. In this way, information about the system that is unavailable from experiments, through MD simulation can be obtained it giving a perception of processes that take place at atomic and molecular level [30], [31]. In addition, in the same way to experiments, MD simulation is possible to set up and control the behavior of environment such as pressure, temperature, and atomic configuration [30]. There are three types of applications where are involved the MD simulation on macro-systems. The first to determine or refine structures through data from experiments. Second to give a description of the system at equilibrium in which is involved the structural, motional properties, and values of thermodynamic parameters. Third to check the real dynamic, that is, developing a system correctly over time [31].

Classically, MD simulation is explained as numerical solving of classical equations of motion for a group of atoms, that is, interaction potential and motion among atoms or molecules are governed by classical Newtonian equation of motion [30], [32]. Here, two assumptions are involved in the integration of the equation of motion that are fundamental in MD simulation. The first assumption is that atoms' behavior is equal to classical entities, that is, they obey Newton's equation of motion, likewise the accuracy of this approach depends of conditions of simulation. Second assumption is related the modeling of how atoms interact among them, for that, it is necessary to have a representative description of those interactions [30].

Atom-atom interactions generate forces which are involved in term of an empirical potential. The forces are related to the first derivatives of the potential with respect to the atom position [32], [33]. A straightforward example of the dynamic motion of a molecular system evaluates from its total energy can be observed in the equation 2.1 according to Hook's law of a small diatomic molecule.

$$E_r = \frac{1}{2}k\left(r_i - r_0\right)^2 \; and \; F(r) = -\frac{dE}{dr} = -k\left(\left(r - r^0\right)\right. \tag{2.1}$$

where E is the energy, F is the force, k is the Hook's constant, r_i and r_0 are the final and initial distances, respectively. The forces are then calculated by the derivative of the energy with respect to the position. Consequently, the Newton law of motion can be employed to determine the motion of molecules or atoms and update the atom positions when it is obtained the forces F(i) on the atom by equation 2.2

$$F(i) = m_i \times a_i \tag{2.2}$$

where m is the atomic mass of each atom and a is the acceleration. Through the calculation of acceleration can be introduce into the equation for the position r$_i$ at time t+Δt, given r$_i$ at time t:

$$r_i\left(t+\Delta t\right)=r_i\left(t\right)+v_i\Delta t+\frac{1}{2}a_i\left(\Delta t\right)^2+... \tag{2.3}$$

Therefore, these last equations lead to generate a trajectory which is an ordered list of 3N atom coordinates for each snapshot at time step in MD simulation. Likewise, the resolution of each equation above mentioned configures a MD basic algorithm as shown Fig. 2.1 [28].

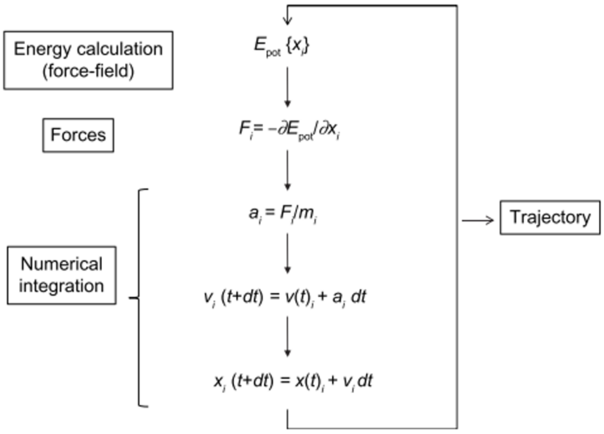

Energy calculation (force-field)
Forces
Numerical integration
$E_{pot}\{x\}$
$F_i = -\partial E_{pot}/\partial x_i$
$a_i = F_i/m_i$
$v_i\,(t+dt) = v(t)_i + a_i\,dt$
$x_i\,(t+dt) = x(t)_i + v_i\,dt$
Trajectory

Figure 2.1. Basic algorithm of MD simulation. The global steps such as energy calculation, estimation of forces, numerical integration and trajectory. E$_{pot}$ is potential energy, t is simulation time, dt is iteration time, F is force component, a is acceleration, m is atom mass, and v is velocity.

2.2.1 Potential energy function

An example of potential energy is the equation 2.1 based on the Hook's law for a diatomic molecule. However, in the biological research the characterization of potential energy is more complex since it determines the stabilities of stable or metastable structures [30], [33]. Moreover, it permits to describe processes such as protein interaction, binding between ligands and proteins, interaction between molecules or behavior of proteins in solvents [30]. Therefore, potential energy is especially deduced from features of the molecular structure and parameters which are obtained by *ab initio*, semi-empirical quantum mechanical calculations and experimental data [30], [33]. This latter concept establishes a force field in which are described general properties of molecules such as torsional barriers, torsional deformation, conformational stereo-isometric energy or to evaluate geometry between interaction molecules or to estimate the vibrational frequency and heat of formation [30]. Although force field equations result complexes by their description, they are easy to solve, but they especially assure to solve energy and forces fast in large systems [28].

Commonly, the potential energy for N atoms at position r_1,...,r_N is described by an empirical force field used for proteins which is composed in two groups of molecular properties. The first group is bonded interaction such as bond lengths, bond angles, torsional angles, and second group is non-bonded interaction that is related with van der Waals interactions and electrostatic contribution, as shown in the equation 2.4 [32], [33]. Then, first derivate with respect to the position of force field equation is related with the force that acting on the atom or molecule of the system [28], [30], [33].

$$E = \sum_{bonds} \frac{a_i}{2}\left(l_i - l_{i0}\right)^2 + \sum_{angles} \frac{b_i}{2}\left(\theta_i - \theta_{i0}\right)^2 + \sum_{torsions} \frac{V_n}{2}\left(1 + \cos\left(n\omega - \gamma\right)\right)^2$$
$$+ \frac{1}{2}\sum_{i=1}^{N}\sum_{i\neq 1}^{N} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{1}{2}\sum_{i=1}^{N}\sum_{i\neq 1}^{N} 4\epsilon_{ij}\frac{q_i q_j}{r_{ij}} \quad where\ r_{ij} = |r_i - r_j| \qquad (2.4)$$

Equation 2.4 shows that the first three terms are related with the internal degrees of freedom in molecules where the term of bond lengths is a harmonic potential between bonded atoms that estimates the energy by the displacement of bond length and term of angles is also a harmonic potential that calculates the energy of adjustment of bond angles in atoms. And the last term of torsions which describes the periodic variation in energy because of bond rotations, here ω is the dihedral angle, γ is the phase shift of the n-fold term and V_n is the barrier height. The last two terms are the nonbonded interactions in which van der Waals interactions are estimated for a Lennard-Jones potential in which there is a repulsive term at very short distances and attractive term accounting for the London dispersion forces between atoms. Second term is electrostatic contribution which is described by the Coulomb electrostatic potential that is attractive or repulsive depending of the effective charge of q_1 and q_2 [30], [32], [33]. Besides, the first four terms are described as short-ranged interactions while the last term is known as long-range interaction [32]. Thus, the total energy is detailed as the sum of bonded and nonbonded interactions as shown in the equation 2.5.

$$E_{total} = E_{bonded} + E_{nonbonded} + E_{other} \tag{2.5}$$

where E_{other} includes the repulsive, van der Waals and Coulombic interaction [30].

2.2.2 AMBER force field

The Assisted Model Building with Energy Refinement (AMBER) is classical molecular mechanics force field. It characterizes the structural and dynamics properties of proteins in water, biomolecules, nucleic acids as well as the study of polymers and small molecules [30]. In contrast with other force field such as CHARMM22, GROMOS or OPLS-AA, AMBER employs torsion parameters that depend of type atoms in the central bond. Therefore, they are improper to keep the stereo-chemistry at chiral centers. Moreover, hydrogen atoms are in a united atom representation, that is, atoms are combined as a heavy atom. Likewise, AMBER includes higher values of hydrogen bond energy than others force field [30]. In the same way to equation 2.4, AMBER force field shares similar properties, as shown in the equation 2.6

$$\begin{aligned} E = \sum_{bonds} K_2 \left(b - b_0\right)^2 + \sum_{angles} \kappa_\theta \left(\theta - \theta_{eq}\right)^2 + \sum_{torsions} \frac{V_n}{2} \left(1 + \cos\left[n\phi - \gamma\right]\right)^2 \\ + \sum_{nonbonded} \frac{A_{ij}}{{R_{ij}}^{12}} - \frac{B_{ij}}{{R_{ij}}^{6}} + \sum_{nonbonded} \frac{q_i q_j}{\epsilon R_{ij}} \end{aligned} \tag{2.6}$$

AMBER has some versions, among them, AMBER-03 [34] which has been employed in this research for all MD simulations. It provides an excellent balance between the extended and helical region distributions and a quantum mechanical calculations for the solvent model [30], [34].

2.2.3 Algorithms for MD simulations

MD simulation employs numerical methods to estimate Newton's equations of motion since an analytical solution is impossible for a huge number of atoms. Verlet integrator is most common algorithm to calculate the trajectories of interacting atoms. It is usually used because it shows minimal local errors and higher stability than others integrators during an MD run, and besides the constrains between atoms are uncomplicated to implement [35]. The Verlet algorithm is derivative by a Taylor expansion about time t of the coordinate of a particle as the next equations

$$r_i\left(t+\Delta t\right)=r_i(t)+v_i(t)\Delta t+\frac{F_i(t)}{2m}\Delta t^2+\frac{\dddot{r}_i}{3!}\Delta t^3+O\left(\Delta t^4\right) \tag{2.7}$$

and

$$r_i\left(t-\Delta t\right)=r_i(t)-v_i(t)\Delta t+\frac{F_i(t)}{2m}\Delta t^2-\frac{\dddot{r}_i}{3!}\Delta t^3+O\left(\Delta t^4\right) \tag{2.8}$$

These equations are added to obtain

$$r_i\left(t+\Delta t\right)\approx 2r_i(t)-r_i\left(t-\Delta t\right)+\frac{F_i(t)}{m}\Delta t^2 \tag{2.9}$$

The equation 2.9 has an error of order $\Delta \mathrm{t}^4$ but it is used to advance the positions of the particle without to use the velocity. However, the estimation of the velocity are based on the equation 2.10 which is derive from of the equation 2.7. For new positions and new forces are calculated.

$$v_i\left(t+\Delta t\right)=v_i(t)+\frac{F_i(t+\Delta t)+F_i(t)}{2m}\Delta t \tag{2.10}$$

Therefore, the new positions are estimated in an iterative cycle where the current positions become the old position [35]. A factor further main is the time-step Δt in any MD simulation. An advantage of Verlet algorithm is that both the positions and velocities can be defined at the same time [36]. The selection of time-step must be a value that algorithm conserves the total energy of the system [35], [36]. In our case, all MD simulations were implemented with this algorithm.

On the other hand, in MD simulation the interactions are divided in short range and long range according to the term in the force field equation, which involves different algorithms to be used in each case. In short range interactions the algorithm employed is Verlet lists which divide the simulation space in small cells with a cut-off distance. Here, each particle is within one cell and the particles that are in the same cell or in an adjacent cell, they should interact in short range. This type of short range contacts belongs van der Waals forces and electrostatic interactions [30]. In the case of long range interactions, the algorithm employed is Particle Mesh Ewald (PME) which uses a fast Fourier transformation in order to calculate the long range contribution [30]. PME has an infinitive range and considers a system as periodic, thus it has a rapid convergence in Fourier space [30].

2.2.4 Periodic boundary conditions

In MD simulation, the system is put into a space-filling box which must have boundary conditions desirable like a crystalline system, that is, for instance a box with particles can be replicated in the space to form an infinite lattice, if one particle leaves from the main box and this one is not replace by another particle of neighbor box leads to errors or problems in the properties of the system commonly called artifacts. For that, it is appropriated to apply periodic boundary conditions that permit minimize edge effects by the translated copies of itself [36], [37]. The space-filling unit cells have several shapes that are used according to the system that will be study or user's needs. The shapes of the box can be cubic, rhombic dodecahedron and truncated octahedron.

Each box has its own properties, for instance in order to study a spherical macromolecule in solution is suggested to use rhombic dodecahedron or truncated octahedron by their shape almost spherical. However, the most common shape is the cubic since it supports any system. The cubic box is given by three vectors (k,l and m) that satisfy conditions such as; (a) $\mathrm{k}_y = \mathrm{k}_z = \mathrm{l}_z = 0$, (b) $\mathrm{k}_x > 0$, $\mathrm{l}_y > 0$, $\mathrm{m}_z > 0$, and (c) $|\mathrm{l}|_x \leq \frac{1}{2}\mathrm{k}_x$,

$|\mathrm{m}|_x \leq \frac{1}{2}\mathrm{k}_x$, $|\mathrm{m}|_y \leq \frac{1}{2}y$. In this way, the inequalities are satisfied by adding and subtracting box vectors and the equality is satisfied by rotating the box [37]. In our systems, it has been employed the cubic box with periodic boundary conditions to perform MD simulations.

2.2.5 Ensembles in MD

MD simulations need to be compatibility with experiments, in order to have a direct comparison between the simulated system and experiment requires that boundary condition will be similar. The thermodynamical boundary conditions are conditioned by a constant temperature and pressure during the MD simulation [38]. For an isolated system with periodic boundary conditions is described by a time-independent Hamiltonian which is invariant in the translational and rotational motions of the system [38]. This system keeps constant the energy E, the total number of atoms N and the volume V leads to generate a trajectory of a microcanonical (NVE) ensemble by solving the Newton's equation. A microcanonical ensemble should be strictly conserved during the simulation, however isolate conditions are not correspond with real experiments since variables such chemical potential μ pressure P or temperature T are not conserved [38]. In an equilibrium simulation, these variables are instantaneous observables which can fluctuate during the MD simulation. Therefore, the ensembles more employed to compare real situations are canonical ensemble (NVT) and isothermal-isobaric ensemble (NPT).

2.2.6 MD at constant temperature

The study of the system at a specific temperature entails to employ a canonical ensemble (NVT). This ensemble requires a thermostat which avoids steady energy drifts by the collection of numerical errors during MD simulation and allows the fluctuation of the temperature over the system. Some strategies have then been designed to build a thermostat.

Berendsen thermostat

This thermostat permits to maintain the temperature of the system through an external heat bath with a fixed temperature T_0 [38]. Here, the rate of change in the temperature is equivalent to the deviation in temperature, as shown in the equation 2.11

$$\frac{dT(t)}{dt} = \frac{1}{\tau}\left(T_0 - T(t)\right) \tag{2.11}$$

where τ is a coupling parameter between the bath and system which is used as an empirical parameter to adjust the strength of the coupling [38]. Moreover, τ conditions the behavior of thermostat since if $\tau \longrightarrow \infty$ is inoperative, then the run is a microcanonical ensemble. In opposite, a small value of τ will produce low temperature fluctuations. Hence, τ should rigorously be chosen as the time step δt. Otherwise, the desire temperature is established through an exponential decay of temperature in the system [38]. From equation 2.11 the change in temperature by each time step is defined as the equation 2.12

$$\Delta T = \frac{\delta t}{\tau}\left(T_0 - T(t)\right) \tag{2.12}$$

Likewise, the velocities are scaled at each time step through a scaling factor 2.13

$$\lambda^2 = 1 + \frac{\delta t}{\tau}\left[\frac{T_0}{T\left(t - \frac{\delta t}{2}\right)} - 1\right] \tag{2.13}$$

Nosé-Hoover thermostat

Nosé-Hoover thermostat considers the heat bath as an integral part of the system, for that, an artificial coordinates and velocities are added whose physical description is a friction parameter ζ, which increases or decreases the acceleration of particles until that the temperature should be equal to the desired value [39]. Hence, the equation of motion is described as the equation 2.14

$$\frac{d\zeta(t)}{dt} = \frac{1}{Q}\left[\sum_{i=1}^{N} m_i \frac{V_i^2}{2} - \frac{3N+1}{2} k_B T\right] \tag{2.14}$$

where Q determines the relaxation of the dynamics friction ζ(t), T is the desire temperature. This equation of motion is implemented by a small modification in the Verlet algorithm to update the position and velocities of the simulation according to the equation 2.15 and 2.16 [39].

$$\zeta\left(t+\delta t\right)=\zeta\left(t+\frac{\delta t}{2}\right)+\frac{\delta t}{2Q}\left[\sum_{i=1}^{N}m_i\frac{v_i\left(t+\frac{\delta t}{2}\right)^2}{2}-\frac{3N+1}{2}k_BT\right] \tag{2.15}$$

$$v_i\left(t+\delta t\right)=\frac{\left[v_i\left(t+\frac{\delta t}{2}\right)^2+\frac{\delta t f_i(t+\delta t)}{2m_i}\right]}{1+\frac{\delta t}{2}\zeta\left(t+\delta t\right)} \tag{2.16}$$

2.2.7 MD at constant pressure

Experiments are normally performed at constant pressure in order to simulate a system with the same condition, it is employed the canonical ensemble (NPT) or also called isothermal-isobaric ensemble. Thus, MD simulation is performed at constant pressure. However, a constant pressure requires a change in the volume of the system [35]. Likewise, as the ensemble NVT, here is necessary to build barostats [40].

Berendsen barostat

The implementation of this barostat allows to rescale the coordinates and box vectors in every step with a matrix μ,hence it is possible to estimate the instantaneous pressure, P which is given by the equation 2.17 [37], [40]. Berendsen barostat generates usually a correct average pressure during the simulation, however there are errors that can be neglected in NVT ensemble [37].

$$P=pT+\frac{\beta}{V}\ \ where\ \beta=\frac{1}{3}\sum_{i>j}f(r_{ij})\cdot r_{ij} \tag{2.17}$$

where β is the virial, V is the system volume, f(r_{ij}) is the force on particle i by particle j. In order to scale the length in the system, it is used a scale factor μ which is given by equation 2.18, where Δt it the integrator time-step, τ_p is the rise time of the barostat, and P_0 is the set point pressure [40].

$$\mu = \left[1 + \frac{\Delta t}{\tau_p}\left(P - P_0\right)\right]^{1/3} \tag{2.18}$$

Parrinello-Rahman barostat

Parrinello-Rahman barostat obtains a exact description of NPT ensemble in the same way to Nosé-Hoover thermostat. This barostat correctly describes small systems which have a high fluctuations in pressure or volume [37]. The box vectors are displayed as matrix b defined in the equation 2.19

$$\frac{db^2}{dt^2} = VW^{-1}b_p^{-1}\left(P - P_{ref}\right) \ \ where \ \left(W^{-1}\right)_{ij} = \frac{4\pi^2\beta_{ij}}{3\tau_p^2 L} \tag{2.19}$$

where V is the box, P and P_{ref} are the current and reference pressures, respectively W^{-1} is an inverse matrix parameter that is associated with the strength of the coupling where β is isothermal compressibilities, τ_p is the pressure time constant and L is the largest box matrix element [37]. The motion of particles with this barostat is defined by the equation 2.20

$$M = b^{-1}\left[b\frac{db_p}{dt} + \frac{db}{dt}b_p\right]b_p^{-1} \tag{2.20}$$

An advantage of this barostat is related with the use in large boxes of simulation since big systems may results in high oscillations, hence it is convenience to employ it. Even though the constant of time will increase [37]. On the other hand, Parrinello-Rahman barostat is not recommendable to high precision thermodynamic calculations [37].

2.2.8 Simulation methods

MD simulation can be developed both homogeneous systems as for instance a box with water molecules with periodic boundary conditions where the simulation is only a few picoseconds and heterogeneous systems where requires longer time of simulations for systems for example as proteins in water [33]. Simulations of nanoseconds (heterogeneous system) are used to characterize ruptures of small groups and determines the dominant contributions to atomic fluctuations [33]. In this vein, MD simulations are fairly similar to real experiments since it is desirable an initial set of atomic coordinates (model system or sample). The coordinates can be obtained from X-ray or by model-building. The set of coordinates of structure is refined using an iterative minimization algorithm to relieve local stresses due to overlap of non-bonded atoms, bond length distortions and others. In this point, the positions and velocities are linked to the system where it must be equilibrated from its initial state [33], [36]. The initial state is usually closed to have a system in the equilibrium. This equilibration state is followed by new velocities which are randomly designated according to Maxwell-Bolrzmann distribution with intervals of dynamical relaxation. A parameter as temperature T of system is estimated by the kinetic energy

$$\frac{1}{2}\sum_{i=1}^{N} m_i \langle v_i^2 \rangle = \frac{3}{2} N k_b T \tag{2.21}$$

where $\langle {v_i}^2 \rangle$ is the average of velocity squared of the ith atom in any Cartesian component, k$_B$ is the Boltzmann constant, and N is the number of atoms in the system. Hence, the equilibrium is gathered through the evolution of trajectories in the time using the Verlet algorithm. However, it must be taken in account the relief of system in the equilibrium because the potential energy will decrease while the kinetic energy will increase, thus the T will also change which will be higher than desired temperature. For avoid that, it is useful to utilize a method known as velocity scaling which warrants to keep a stable T until the final equilibration system (see equation 2.21) [36]. For all MD simulations in this study has been performed through this method. Consequently, finished the equilibrium stage, the main part of MD simulation will start. At this point forces, positions and velocities are estimated for each particle at each time-step again, besides these positions and velocities are recorded until the simulation ends, although the velocity scaling is not applied [36]. The outcomes of MD simulation are positions and velocities of each particles of the system at equilibrium which permit to quantify properties of physical interest such as average of kinetic or potential energy in the full simulation run [36].

2.3 Docking

Molecular docking permits to predict of the structure of ligand-receptor complexes with the best matching from computational approach [41], [42]. Three elements are key in the molecular docking which are (a) systems (ligand-receptor) (b) conformational space and (c) ranking of potential solutions [41]. Both a receptor or a ligand can be a protein, ligand or small molecule [42]. The exploration of molecular docking in protein-ligand complexes have specially been successful permitting to create database of screening in drug discovery [42]. The develop of complexes have permitted to determine roles and functions in interaction's ligand-receptor. In the case of protein-protein complexes is still more complex because their structures have a biological function. Therefore, the accuracy in the prediction of docked complexes are substantial to obtain functional information about this system [43]. However, any docking programs have two main limitations which are the conformational degrees of freedom or a correct docked orientation with a great likelihood and the scoring function in order to discriminate between a correct or incorrect docked orientation [41], [43]. Methods usually used to generate a docked complex by protein-protein are based on the shape and chemical complementarity [41], [42], [43]. The shape complementarity is most robust score function based on the surface shape of each protein to form the complex. For this purpose, surfacing algorithms are used which estimate the solvent-accessible or the occupation of spaces in cells generated by a grid in the protein space [43]. Likewise, the algorithms explore the global energy minimum by means of potential energy surface using two approximations such as rigid docking or flexible ligand [41], [42]. In a rigid docking, the ligand examines different position using translational and rotational degrees of freedom, however there is a limitation based on the influence of the ligand over receptor since by induced-fit. In the case of flexible ligand, it is added a torsional degrees of freedom of the ligand [41], [42].

2.3.1 Docking procedures

Fig 2.2 [43] shows the procedure of molecular docking where the coordinates of two molecules must be entered in order to predict a docked complex.

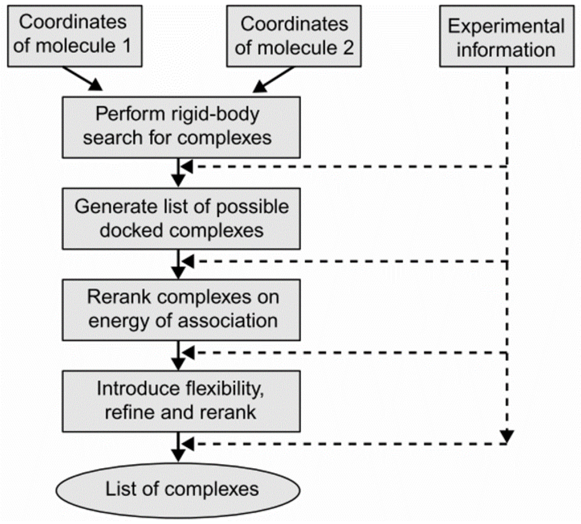

Coordinates of molecule 1
Coordinates of molecule 2
Experimental information
Perform rigid-body search for complexes
Generate list of possible docked complexes
Rerank complexes on energy of association
Introduce flexibility, refine and rerank
List of complexes

Figure 2.2. Procedure of docking.

Searching of first docked complexes

In this procedure, proteins are treated as rigid bodies in order to identify a set of candidates structures using a scoring function [43]. The implicit method employed in this step is the Fast Fourier Transform (FFT) which divides the surface of proteins into a voxel grid. Using this approach has been provided a scoring function which can estimate all relative translations and rotational motions searching the favorable orientation of the molecules, According to this approach allows to give an approximation of electrostatic energies and penalize the overlap between the cores of molecules. Besides, it has been developed other scoring functions in which are used physicochemical features of the proteins such as surface area, complementarity of curvature, and penalize the overlap of core protein or the use of Fourier correlations to simplify the problem of estimation of the complementarity between surfaces in distinct orientations [41], [43].

Rescoring

The new structures generated by a global searching are re-ranked in order to obtain near-native orientations through an complex scoring function [43]. One method is usually using the statistic of residues contacts in the docked complexes. Likewise, other methods add terms such as electrostatics interactions, hydrogen bonding interactions, or charges in the buried residues [43].

Flexibility on the structure

In this stage one model is usually characterized through the movement of side chains or backbone since rearrangements of side chains for their combination of rotamers may dictate the docking configuration. It is minimized by an energy function which is applied by a molecular mechanics force field for proteins [43]. Although, an energy function may be irregular for each model, it is usually employed as an efficient optimization algorithm. Additionally, extra biological information can be employed to constrict docked complexes. For instance interface residues of proteins or mutagenic information reducing the time and simplify the search of docked complex [43].

2.3.2 Algorithms use in docking

The algorithms are related with the flexibility during docking process since there are different conformations when ligand-receptor interacts between them. These algorithms are divided in three types such as systematic, stochastic and deterministic searches. Although some algorithms can combine more than one of these approaches [41], [43].

Systematic search algorithms

These types of algorithms are focused on each formal degree of freedom of ligand which is divide rigid and flexible regions by a grid. Each degrees of freedom of ligand is directly related with the number of evaluations. The coupling between ligand and receptor is in the active site through a systematic scanning of torsion angles [41]. This procedure has a limitation corresponding to small ligands since numerous positions on the surface of receptor can interact with the ligand generating an increase the number of docked structures without a correct position [41], [42].

Stochastic search algorithms

These algorithms allow to chance one degree of freedom of the system in a time step. However, a limitation is an uncertain convergence, hence it is performed multiple runs in order to overcome this limitation. The highlight examples of these algorithm are Monte Carlo (MC) methods and evolutionary algorithms. Regarding to MC, the ligands make random changes in the translation and rotation motions as well as in the torsion angles. For each move, the energy of the new arrangement is minimized according to the Metropolis criterion. Likewise, the number of cycles to gather a global energy minimum can be large which is related to a change of the temperature. Otherwise, evolutionary algorithm employs several parameters such as mutations rate, crossover rates, number of evolutionary rounds and size of population that calculate conformational changes in the structure of docked complexes. Likewise, this algorithm classifies the structures as fittest individuals according to global energy minimum that are used several times to improve the final docked complex, this process is known as crossover of generations. Besides, it is possible to make mutations, that is, changes in the structure getting genetic diversity (more structures) and avoid a rapid convergence [41], [42].

Deterministic search algorithms

These algorithms are associated with an initial state in which is introduced a move to produce a next state, that is, a new docked complex which must be similar or lower in energy than an initial state. The limitation with these types of algorithms are often confined in local minimum since the energy barriers are high to cross. Energy minimization and MD simulation are examples of deterministic search algorithms. MD simulation approach has been used to examine the binding free energy landscape or the potential energy surface of all degrees of freedom of the ligand-receptor, however the main limitation is about the time of simulation in each complex [41], [42].

2.3.3 Protein-protein docking

Docking between proteins have been considered as essential element in the understanding of cellular pathway, macromolecular interactions and the inhibitor design [41]. Nevertheless, protein-protein docking is a great challenge in comparison with other types of docking where are related with small molecules. The coupling between two proteins is difficult task because the number of degrees of freedom is huge as well as the sizes of molecules [41]. Therefore, it has been designed specifically algorithms and scoring functions for these types of docked complexes, although they share the same principles as it was described above [42]. Within the framework proposed in the Fig. 2.2, the standard procedure in protein-protein docking algorithms is supported by the use of rigid body approximation since there is a large number of degrees of freedom and the interactions sites are arduous to predict in protein-protein docking [42]. Likewise, the scoring functions are less strict because hiding atom clashes even at near-native configurations. The variability is given by the protein surfaces and when these are sampled, it is created a huge number of complexes which are ranked by scoring functions depended on geometric filter which eliminates configurations that are unacceptable the complex structure [42]. Generally, the outcomes obtained by protein-protein docking algorithm are satisfactory to build known complexes, however in complex structures without known bound structure previously, the rearrangement will depend of input data coordinates and the native complex structures [41], [42]. Although, the quality of prediction will diminish. The prediction of docked complex depends of the extension in the rearrangement because the surface of molecules is in constant motion which are associated with the side chains in protein structure. Hence, several approaches have been involved in the docking algorithms to attend extend systems. The first is based on the Cα backbone atoms instead of complete description of side chains, thus the motion at the surface atoms are only taken into account. However, it is necessary to apply improved scoring schemes [41]. Second approach is based on the side chain motions, however the large backbone motions are not involved limiting the solutions to obtain complexes [41]. Third approach is known as hinge-bending motions where ligands undergo translations and rotations in order to dock to the surface of the receptor [41]. An advantage of this approach is nonessential to know the binding sites or the hinge locations with respect to the receptor. Therefore, in order to bind ligand-receptor is used descriptor points in the surface of both proteins, where if each protein has similar point, that is, a configuration equal it is considered as a hinge location which is ranked. The hinge location with a high score performs a coupling with the receptor [41].

2.4 Gibbs binding free energy

When a docked complex is formed there is a complementarity of the shape between receptor and ligand as a lock-and-key which is known as molecular recognition model [42]. This molecular recognition model shows physical-chemistry properties such as van der Waals interactions, electrostatic interaction, hydrogen bond interaction together with hydrophobic effect, and entropy attractions which influence the stability of a complex. According to the equation 2.22 can be measured the stability of complex by estimating of equilibrium binding constant K_{eq} that allows to related K_{eq} with the Gibbs binding free energy ΔG_{Bind}. Likewise, the stability can be estimated by on-rate k_{on} and off-rate k_{off} constants of the reaction.

$$\Delta G_{bind} = -RTlnK_{eq} = -RTln\left(\frac{k_{on}}{k_{off}}\right) \tag{2.22}$$

The constant K_{eq} depends directly of variables such as temperature, pH, pressure, ionic strength and concentration of solutes [42]. Although this constant involves variables that influence in the experiments, the comparison between theoretical and experimental values of Gibbs binding free energy should be taken with attention [42]. However, in the estimation of stability in complex, the Gibbs binding free energy can be separated in two contributions as enthalpic and entropic in order to calculate the free energy of the complex, ligand and receptor as shown in the equation 2.23

$$\Delta G_{bind} = \Delta H - T\Delta S = G_{complex} - (G_{receptor} + G_{ligand}) \tag{2.23}$$

2.4.1 Methods of estimation for Gibbs binding free energy

There are three ways to estimate the Gibbs binding free energy ΔG_{Bind} which are usually more precise, that are Thermodynamic Integration (TI), Free Energy Perturbation (FEP) and molecular mechanics force field [42].

Thermodynamic Integration (TI)

TI is the most common approach in the estimation of ΔG_{Bind} in the equilibrium. This method has a scaling parameter λ since it generates equilibrium ensembles of configuration with multiple values of λ. TI can calculate an accuracy value of ΔG_{Bind}, however it is usually computationally expensive because each λ values should be equilibrated. ΔG_{Bind} is found according to the equation 2.24 [44].

$$\Delta G_{Bind} = \int_{\lambda=0}^{1} d\lambda \left\langle \frac{\partial U_{\lambda}(x)}{\partial \lambda} \right\rangle_{\lambda} \tag{2.24}$$

where the intergral shows an ensemble average in each values of λ, while $U_{\lambda}(x)$ is the functional form that depends on the scaling methodology. In addition, λ values can be simulated a finite number which involves errors in the equilibrium sampling in each λ, hence the integral must be approximated by a sum [44].

Free Energy Perturbation (FEP)

FEP approach employs the λ as TI in each equilibrium simulation, however it uses a exponential average between neighbor λ values to estimate the difference in ΔG_{Bind}. These differences are summed to get the total free energy difference of ΔG_{Bind}. Likewise as TI, there is a limitation based on spacing of λ values which must be small to reach an overlap between the configurations spaces corresponding to λ_i or λ_{i+1} [44]. ΔG_{Bind} is calculated according to the equations 2.25 and 2.26 which are approximated for λ values between 0 and 1 using a forward and reverse estimation [44].

$$\Delta G_{Bind} = -k_B T \sum_{i=0}^{n-1} ln \left\langle e^{-\beta\left(U_{\lambda_{i+1}}(X_i) - U_{\lambda_i}(X_i)\right)} \right\rangle_{\lambda_i} \tag{2.25}$$

$$\Delta G_{Bind} = +k_B T \sum_{i=0}^{n-1} ln \left\langle e^{-\beta\left(U_{\lambda_i}(X_i) - U_{\lambda_i}(X_{i+1})\right)} \right\rangle_{\lambda_{i+1}} \tag{2.26}$$

Molecular mechanics force field

This method is used to calculate the $\Delta \mathrm{G}_{Bind}$ through a force field that only estimates the enthapic contribution, although the entropic contribution is estimated both solute and solvent by different methods [42]. In this method is normally calculated the initial and final states to obtain $\Delta \mathrm{G}_{Bind}$ which is the main difference with respect TI and FEP methods since they consider intermediate states. This method employs a force field as the equation 2.4. The first three terms are used to estimate the internal energy through corrections from ideal values of bond lengths, bond and dihedral angles [42]. The four term involves the enthalpic contribution according to the Lennard-Jones potential [42]. The entropic effect for solutes is estimated through statistical mechanics approach, quasi-harmonic analysis or statistical thermodynamics, but for the solvent effects is estimated in two terms nonpolar and polar. The nonpolar term is calculated using the surface area of the solute [42]. The polar term is estimated by the differences in electrostatic energy between solute embedded in a low and high dielectric medium. Besides, this term can be calculated with Poisson-Boltzmann equation with a continuum solvent approximation [42]. Therefore, these terms together allow to determine the strength of the interaction between proteins and ligands [42]. The force field equation 2.4 can be inserted in a general equation in order to obtain the Gibbs binding free energy as shown in the equation 2.27

$$\Delta G_{bind} = E - T\Delta S_{solute} + \Delta G_{solvent} \tag{2.27}$$

Therefore, the solute entropy in the second term involves the translational, rotational, conformational and vibrational entropy and third term consists in nonpolar and polar terms, where is also taken into account the enthalpic and entropic effects in the solvent [42]. Besides, this equation has been employed in different systems where the Gibbs binding free energy is calculated to different structures through MD simulations with approaches such as molecular mechanics energies combined with the Poisson–Boltzmann (MM/PBSA) or generalized Born and surface area continuum solvation (MM/GBSA) methods [42].

2.4.2 Time-consuming process of ΔG_{Bind}

In order to calculate the Gibbs binding free energies require a time-consuming process. The different methods such as FEP or TI require a high computational cost and large time of simulation. On the other hand, MM/PBSA method has been used by its small computational cost, but it is limited to screening of a short number of ligands and small systems. It has been established methods that are fast and accurate to obtain the ΔG. These methods are namely first-principles methods, semiempirical method, and empirical methods [42].

First-principles methods

This method is applied in several programs where it is estimated a force field function that is implemented the van der Waals and Coulomb terms as well as the electrostatic interaction of the solvent. However, entropic terms and ligand energies are ignored to facility the estimation of ΔG, hence it may be overestimated the complex stability [42].

Semiempirical methods

These methods are less computational cost since they take the initial an final states to be sampled in order to estimate the ΔG. However, these types of methods perform several approximations that reduce the time to estimate the ΔG. Several programs employ this approach, thus the methodologies will be different [42].

Empirical methods

In contrast with other methods, the empirical approach has some limitation in their functions such as the accurate of each term and how these terms were evaluated. Besides, the prediction of ΔG is successful whether the systems are similar to the set of structures which are before evaluated, that is, it can exist troubles with the transferability of methods with other systems. Variables such as pH, salt concentration and temperature which are involved in the constant that are used in the functions to estimate ΔG. However, the time of process is short in comparison with other methods [42].

Chapter 3

Methods

3.1 Sequence analysis of NS5 and STAT2

The amino acid sequence of the target proteins, NS5 (UnitProt code: B1P6I2) and STAT2 (UnitProt code: P52630) were extracted from the UniProt protein sequence database ($http : //www.uniprot.org$) each sequence contained 902 and 851 residues, respectively. Each sequence target protein was analyzed to determine the amino acid composition, disordered regions, binding residues, contact sites and secondary structure through the programs MEGA6 [45], DISOPRED [46], RaptorX binding [47], RaptorX contact prediction [48], RaptorX property prediction [49], and GOR4 [50]. MEGA is implemented to develop comparative analysis of DNA and protein sequences [45]. DISOPRED identifies residues which are likely to be natively unfolded [46]. RaptorX binding predicts the binding sites based on a sequence of a protein [47]. RaptorX contact predicts the contact map of a protein sequence without using any templates. RaptorX property is used to predict the secondary structure, solvent accessibility, and disordered regions without using templates [49]. GOR4 infers the secondary structures in proteins sequence [50].

3.2 Molecular modeling of NS5 and STAT2

The NS5 three-dimensional model (PDB code: 5TMH) was obtained from protein data bank ($https://www.rcsb.org$). However, only two fragments (PDB codes: 5OEN and 2KA4) has been reported for STAT2. Three-dimension models were generated through algorithms of protein threading and homology alignment programs known as I-TASSER [51] and Phyre2 [52]. I-TASSER identifies template proteins from structure databases that have similar structure to the query protein sequence by position-specific iterated BLAST (PSI-BLAST). In this way, it creates a sequence profile which is used to predict the secondary structure using PSIPRED [53]. Templates with similar folds are retrieved to build an assembly, whereas for the structures not found is employed *ab initio* modeling to complete a three-dimensional structure. Then, the models are reassembled into full-length models. Moreover, they are optimized in their H-bonding to avoid steric clashes in full atomic models. The best model was selected through the confidence score (C-score) which has a range from -5 to +2 where a score close to -5 indicates correct model topology [51]. Phyre2 follows a procedure similar to I-TASSER, however it uses a hidden Markov model (HMM) which predicts a secondary structure [54]. This profile is scanned in a precompiled database of HMMs Fold library generating a crude backbone model. Subsequently, the crude model is fitted with fragments through a loop modeling. Fitted fragments are scored by empirical energy parameters. Side chains are then fitting to the backbone by R3 protocol which uses a fast graph technique and side chain rotamer library [52]. The best model was selected in terms of the highest level of confidence and identity percentage.

3.3 Structures quality check and structural alignment

In order to evaluate the NS5 and STAT2 models quality were used the programs Verify-3D [55], ERRAT [56], PROCHECK [56], and VADAR 1.8 [57]. As NS5 is provided from a crystallographic structure, the parameters obtained were employed as reference to STAT2.

Verify-3D is used to resolve the relationship of an atomic model with its own amino acid sequence and comparing the outcomes with highly refined structures [55]. ERRAT was used to determine the overall quality between model-building and refined structures [56]. PROCHECK was used to check the stereo-chemical quality of the protein structure by overall geometry features [56]. VADAR 1.8 was used to analyze protein structures through extensive comparison to published data and careful visual inspection [57]. STAT2 experimental fragments reported in the database were employed to perform a structural alignment with the STAT2's model obtained by protein threading alignment through TM-align [58]. TM-align uses an algorithm that compares two independent proteins structures of unknown equivalence. The superposition measure between two structures employ the TM-score which has a value between 0 to 1, where 1 suggests a perfect splice between two structures [58].

3.4 Molecular dynamics simulation of NS5 and STAT2

GROMACS 5.1.2 [37] was used to perform the MD simulation of NS5 and STAT2 as well as NS5-STAT2 complex. MD simulations were conducted by using AMBER-03 force field [34]. The systems were established using the following features; cubic boxes filled with SPC216 water molecules, TIP3P water [59], Na^+ and Cl^- counter ions were added to neutralize the system with a concentration of 0.1 M in order to mimic physiological conditions of the cells and periodic boundary conditions. PME was used for non-bonded interactions such as electrostatic interaction and van der Waals with a cut-off of 12Å and a 2fs time step during the simulation. The energy minimization was obtained through steepest-descent algorithm and the maximum force of the system was set to 100 kJ·(mol·nm)$^{-1}$ on any atom.

NVT and NPT ensembles were equilibrated using Berendsen thermostat [37],[40] and Nose-Hoover thermostat [37], [60] for 500 ps. Parrinello-Rahman barostat [37], [61] was employed to maintain the pressure isotropically with a value of 1.0 bars and a compressibility of $4.5x10^{-5}bar^{-1}$. The systems were subjected to 50 ns of production which was initialized using output data retrieved from previously run equilibration simulation at 310 K and 1 atm. Besides, all bonds length containing hydrogen were constrained using the Linear Constraint Solver (LINCS) algorithm [62]. Gromacs utilities was used to analyze MD trajectory and the charts were plotted using Grace ($http://plasma-gate.weizmann.ac.il/Grace/$).

3.5 NS5 and STAT2 ensembles for protein-protein docking

In order to explore the structural conformations generated from MD trajectories of 50 ns both NS5 and STAT2, a clustering was made. The clustering is based on Root Mean Square Distance (RMSD) of Cα atoms with a cut-off of 2.2Å for each trajectory through the GROMOS clustering algorithm [63] which is executed in the *gmx cluster* tool of GROMACS 5.1.2. A representative structure of each protein's trajectory was extracted in order to be used in an ensemble docking. This representative structure for each cluster (NS5 and STAT2) was also selected in terms of the lowest energy and stereo-chemical quality for the docking ensemble. Protein-protein docking was performed using ClusPro 2.0 [64] and Pydock [65]. ClusPro docking server uses docking approach known PIPER'S efficient FFT which generates complexes considering scoring functions namely as balance, electrostatic, hydrophobic, and van der Waals interactions. In order to obtain the binary complex, the first ten structures were selected according to higher populated cluster in the scoring function of balanced interaction [64]. Pydock docking server employs a Fast Fourier Dock (FTDock) which use a FFTW 2.1.5 library to generate docking poses by a global scanning of translational and rotational space, followed by optimization. Then, the models are scored by Pydock scoring algorithm which has an efficient empirical potential, composed of electrostatic and desolvation terms and limited contribution of van der Waals energy [65]. In the same way to ClusPro, the first ten models with the high stabilization energy were then chosen.

3.6 Protein-protein interaction and Binding energy of binary complex NS5-STAT2

Each binary complex selected from ClusPro and Pydock was quantified by the strength of the protein-protein interface and interaction energy through the programs PPCheck [66] ($http://caps.ncbs.res.in/ppcheck/$), and FoldX [67]. Moreover, with the software PISA [68] ($http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html$) was calculated the interface area and the solvation free energy gain upon formation of the interface.

PPCheck employs pseudo-energies as van der Waals, electrostatic and hydrogen bond interactions to calculate the strength of these non-bonded interactions for a protein-protein complex. Hydrogen bond energy is computed using the equation 3.1

$$E_{H-bond} = q_1 q_2 \cdot \left[\frac{1}{r(ON)} + \frac{1}{r(CH)} - \frac{1}{r(OH)} - \frac{1}{r(CN)} \right] \cdot 332 * 4.184 kJ \cdot mol^{-1} \quad (3.1)$$

where q_1 and q_2 are partial atomic charges, r is the inter-atomic distance between neighbor atoms. The van der Waals interactions are obtained using the equation 3.2

$$E_{vdW} = 4.184(E_i E_j) \times \left[\left(\frac{R_i + R_j}{r} \right)^{12} - 2 \left(\frac{R_i + R_j}{r} \right)^{6} \right] kJ \cdot mol^{-1} \quad (3.2)$$

where R is the van der Waals radius for an atom, r is the distance between atoms, and E is the van der Waals well depth. The electrostatic interaction are estimated based on the Coulomb's law using the equation 3.3

$$E_{el} = \frac{4.184 * 332 - q_1 q_2}{D \cdot r} kJ \cdot mol^{-1} \quad (3.3)$$

where q_1 and q_2 are also partial atomic charges, r is the distance between atoms and D is the diaelectric constant of surrounding. All these interaction energies are summed in order to obtain the total stabilizing energy which is divided by the total number of interface residues to gather the energy per residue in the complex. The latter one has a range between -2kJ$\cdot$mol^{-1} to -6kJ$\cdot$mol^{-1} with a number of residues between 51 to 150 at the interface. Therefore, a binary complex with a value close to -6kJ·mol^{-1} is considered as a stable interface and correct docking pose [66].

FoldX uses a empirical force field that it has been used to calculate the Gibbs binding free energy (ΔG_{bind}) which is related with the thermodynamic dissociation constant (K_d) according to the equation 3.4 in order to determine the interaction between two molecules.

$$\Delta G_{Bind} = -RTln(K_d) \tag{3.4}$$

where R is the gas constant and T is the temperature in K. The ΔG_{bind} of a complex (AB) is given by equation 3.5;

$$\Delta G_{Bind} = \Delta G_{AB} - (\Delta G_A + \Delta G_B) \tag{3.5}$$

Therefore, FoldX estimates the change in Gibbs energies of the complex (ΔG_{AB}) and of two molecules A and B alone [67]. For that, FoldX utilizes a force field known as FoldX force field (FOLDEF) to describe the energetic contributions in empirical terms to calculate the free energy (kcal·mol^{-1}) of stabilization. Each term of equation 3.6 has a constants ($a..l$) related to the weights of the different energy terms [67].

$$\begin{multline}\Delta G = a \cdot \Delta G_{vdw} + b \cdot \Delta G_{solvH} + c \cdot \Delta G_{solvP} + d \cdot \Delta G_{wb} \\ + e \cdot \Delta G_{hbond} + f \cdot \Delta G_{el} + g \cdot \Delta G_{kon} + h \cdot T \cdot \Delta S_{mc} \\ + k \cdot \Delta S_{sc} + l \cdot \Delta G_{clash} \tag{3.6}\end{multline}$$

In order to calculate the Gibbs free energy, the hydrophobic ΔG_{solvH} and polar ΔG_{solvH} terms are contributions in the interaction with the bulk solvent where b and c have been obtained by the transfer of amino acid from water to an inorganic solvent which mimics the transition from folded to unfolded state in hydrophobic environment in a native state. ΔG_{wb} term is related with the water molecules that are persistent in the interaction with protein groups that making more than two hydrogen bonds. ΔG_{vdw} term is calculated in the same way to the hydrophobic and polar terms but experimental energies from water to vapor are considered. ΔG_{hbond} term is calculated based on simple geometric considerations and their energy. ΔG_{el} term is estimated by a simple implementation of Coulomb's law, here, hypothetical atoms are included to Coulombic interaction in order to computed specific aspects such as; helix dipole interaction. The dialectric constant used is scaled with the burial of the bond. ΔG_{kon} term is electrostatic contribution for protein complex which is estimated between atoms of different polypeptide chains. ΔS_{mc} entropy term computes the fixing of backbone derived by statistical analysis of the phi-psi distribution of a amino acid. This term is estimated by the accessibility of the main chain atoms and hydrogen bond interactions in relation to residue or its direct neighbors. ΔS_{sc} term is related to fix a side chain which is computed by a set of entropy terms to the burial of the side chain. Finally, ΔG_{clash} term is calculated by the steric overlaps between atoms in the structure [67].

According to the results obtained through PPCheck analyses, the binary complexes with higher total stabilizing energy were selected to perform MD simulations of 50 ns with the same operational parameters discussed for the case of NS5 and STAT2. Besides, analysis of protein-protein interactions in the binary complexes using Protein Interaction Calculator (PIC) web server (*http://pic.mbu.iisc.ernet.in/*) [69] and the contact map using COCOMAPS (*https://www.molnac.unisa.it/BioTools/cocomaps*) [70] which recognizes the interaction surface were added in order to reconfirm the binding sites interface in binary complex of NS5 and STAT2 [25].

3.7 Electrostatic Calculations

The lowest energy structure of NS5, STAT2 and NS5-STAT2 complexes were used to perform the electrostatic calculations. The Poisson-Boltzmann (PB) equation implemented in the APBS Program [71] as plug-in added in the program PyMOL 2.2.2 (*https://pymol.org/2/*) was used. Each atomic coordinates were prepared using the method pdb2pqr [72] which adds hydrogens, missing sidechain atoms and assigns the proper atomic radii and charged using AMBER force field. The parameters implemented in the analysis by default in the software were; ionic radius of 1.5Å, dielectric constants inside protein 2 and for water 78 and ionic strength 50mM. Single Debye-Hückel was applied for the dialectric boundary calculations [73].

Chapter 4

Results and discussion

4.1 Analysis of amino acid sequences

In this first section was performed analyses based on the amino acid sequences in order to achieve a general view of the structural features of NS5 and STAT2 which were besides compared with its three-dimensional structures and behavior by MD simulations.

NS5 and STAT2 sequences were analyzed in their amino acid composition in order to determine the presence of residues such as Ser, Thr, and Tyr that are relevant in the process of phosphorylation. These residues were presented in both sequences with 6.61%, 4.96% and 2.13% for STAT2 and 5.65%, 5.99% and 3.22% for NS5, respectively (see Fig. 4.1). The process of phosphorylation has been associated with conformational changes in the protein affecting its function [18]. Particularly, STAT2 is known that needs to be in the phosphorylated form to regulate the viral response pathway [16]. On the other hand, the interaction of NS5 with other proteins has been observed that diminish in its hyperphosphorylated form [19]. The presence of these types of residues are relevant in both sequences, however during the process of MD simulation the non-phosphorylated forms were used.

Disordered regions in the proteins play a role in their functions such as protein-protein interaction networks, cellular differentiation, human diseases or others [46]. For instance, an intrinsic disorder of a protein allows the binding with multiple targets as well as increasing the efficiency of binding since a disordered region provides flexibility to the domains improving the movement and enrollment with binding partners [74], [75]. Hence, a strict inspection of disordered residues in the targets sequences was performed where it was found that STAT2 has 196 disordered residues which are distributed in the regions of 130-141, 185, 326, 371-373, 375-394, 410-423, and 706-851. While, NS5 has barely 23 disordered residues which are found in the regions 1-6, 269-278, 460-466 of sequence.

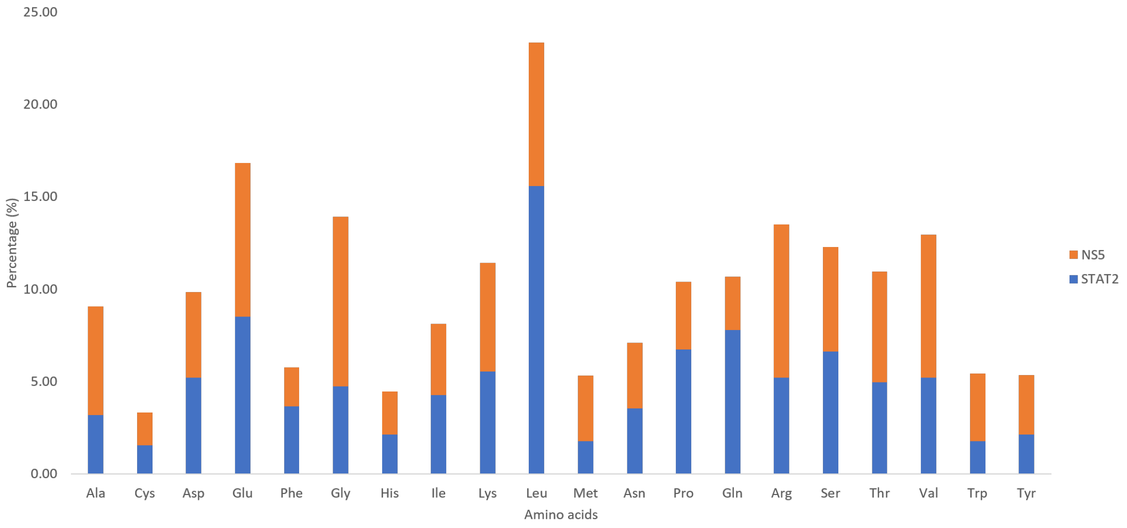


Figure 4.1.NS5 and STAT2 amino acid composition based on their sequence.

The difference in the amount of disordered regions between NS5 and STAT2 show that STAT2 has a higher level of disordered in its structure than NS5 (see Fig. 4.2). These results have been consisted with other studies where at least 30 residues in length form a disordered region in mammalian signaling proteins, and more amount of them is classified as a long region of disordered [74]. On the other hand, within the disordered residues, there are 21 and 43 residues which have been considered as binding residues both to NS5 and STAT2, respectively (see Fig. 4.2). These binding residues in the disordered regions have shown to keep different properties in protein-protein interactions in comparison with residues of ordered regions [76]. The hydrophobic interactions are more favored with disordered residues than their remaining amino acids. Indeed, hydrophobic interactions are stronger than the polar ones which produces more intermolecular contacts, and suggests a better fit with their counterparts through disordered residues [76]. Another property associated with disordered proteins is the binding free energy which is used to assist the protein folding in those zones, that is, with more disordered segments the folding could be energetically demanding [76].

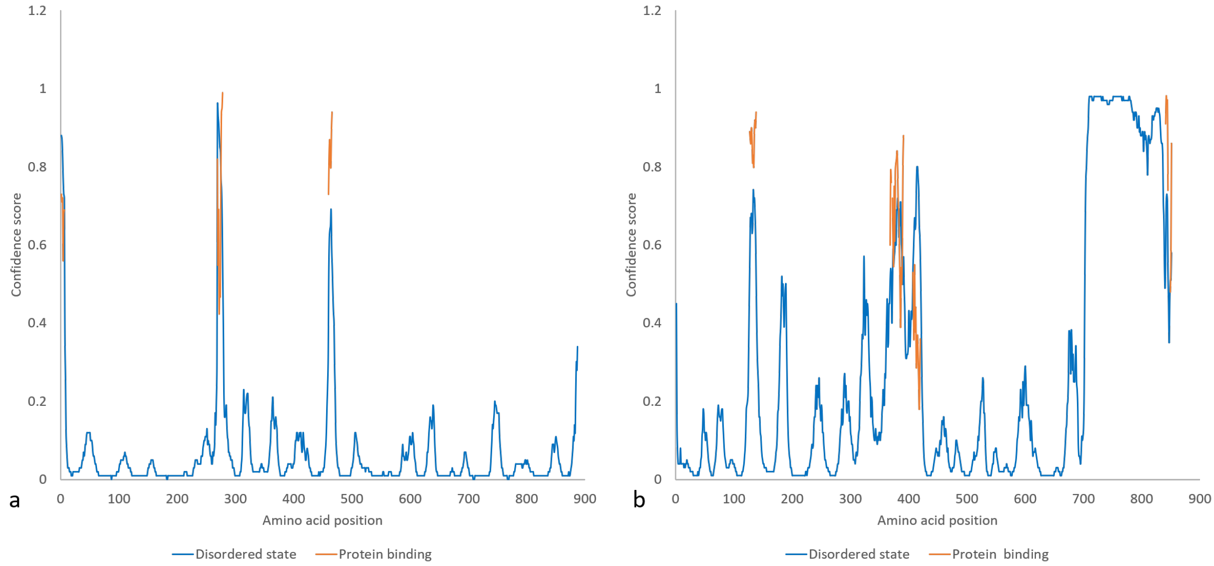


Figure 4.2. Analysis of disordered regions and binding residues in target sequence proteins through DISOPRED. (a) NS5 shows disordered regions in 1-6, 269-278, 460-466 and (b) STAT2 in 130-141, 185, 326, 371-373, 375-394, 410-423, and 706-851 of the sequence. Peaks (blue) over 0.5 confident score are considered regions with high level of disordered. Peaks (orange) are regions with binding residues.

The role of each amino acid is to stabilize the final folding of a protein, for which they generate a local conformation called motives that describe the secondary structure [77]. The presence of motives for NS5 and STAT2 amino acid sequences have been predicted through RaptorX and GOR4, as shown in the table 4.1. These values show that both sequences share common structural motives corresponding to α-helix, coiled-coils and β-sheet conformation. The trend shows that α-Helix and coiled-coil conformations have a high prevalence, in contrast to β-sheet which presents the lowest percentage. The determination of the secondary structure is a way to predict the three-dimensional structure [77], [78]. Likewise, the compactness of the chains in the proteins is associated to an elevated percentage of secondary structure [77]. A maximum compact chains has also been related to a ratio of α-helices and β-sheets roughly equal, that is, similar amount of secondary structures as a real protein [77].

The prediction of residue solvent accessibility aids to elucidate the relationship between sequence and structure because it plays a role in the spatial arrangement and packing of the protein, for instance the protein folding which directly has correlation between the hydrophobic forces and exposure of buried residues [79], [80]. Likewise, it was found that solvent accessibility has a role in the protein's function by sites of protein hydration [81]. Regarding to sites of protein-protein interaction, as well as, active sites are directly influenced by residues with a strong solvent accessibility [80]. The table 4.1 shows solvent access prediction for NS5 and STAT2. In contrast to NS5, STAT2 displays a higher solvent access in the medium residues. However, the number of buried residues in NS5 is higher than STAT2, thus there is a low accessibility to the solvent. This high percentage of buried residues in NS5 suggesting that its stability as protein could be altered easier than STAT2 since many studies have demonstrated that the stability may be affected by change of hydrophobic residues in the buried zones [82]. In the case of exposed residues for NS5 and STAT2 have a similar behavior because the number of residues is almost equal. Hence, as above was mentioned an idea of shape could be provided, consequently NS5 could be consider as globular protein since those types of proteins generally tend to acquire hydrophobic residues instead hydrophilic residues in the interior of the protein while that the hydrophilic residues are exposed to the solvent [79]. In this sense, STAT2 according to the values observed invites to think in a combination between globular and a fibrous protein.

Table 4.1. Prediction of secondary structure and solvent access to the amino acid sequence of NS5 and STAT2 through RaptorX property prediction and GOR4.

Software	Property	STAT2	NS5
RaptorX	α-Helix (%)	38.00	36.00
	β-sheet (%)	11.00	12.00
	Coiled-coils (%)	50.00	50.00
GOR4	α-Helix (%)	36.95	34.81
	β-sheet (%)	15.47	22.51
	Coiled-coils (%)	47.58	42.68
RaptorX	Exposed (%)	23.00	28.00
	Medium (%)	52.00	32.00
	Buried (%)	23.00	39.00

4.2 Molecular modeling and validation of STAT2

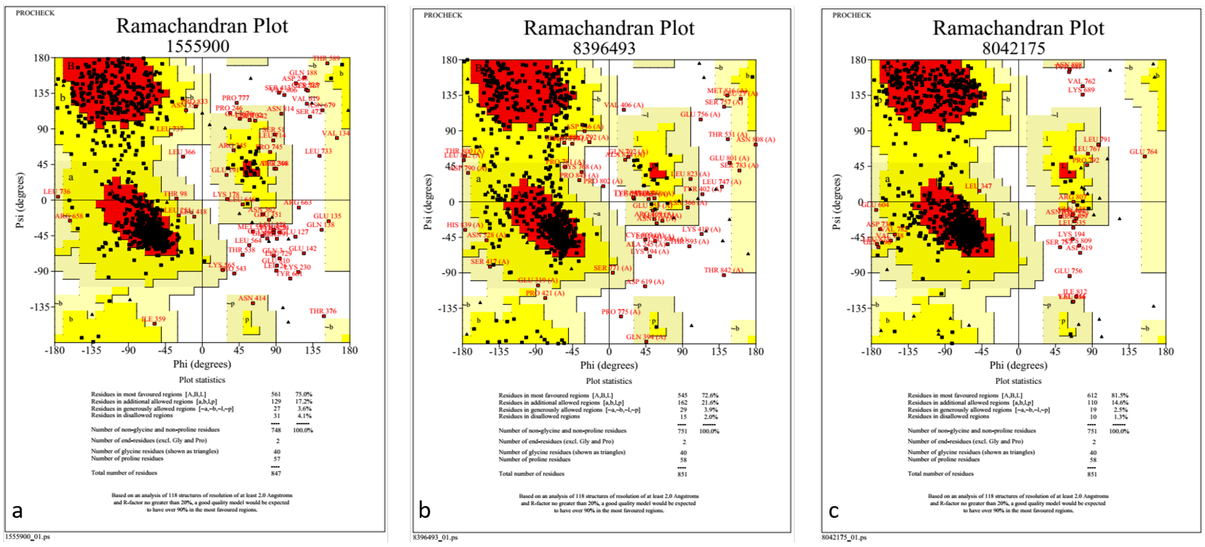
The three-dimensional model of STAT2 was generated using the amino acid sequence by two different approaches namely protein threading of I-TASSER and homology of Phyre2 [51], [52]. Through I-TASSER was obtained five models. The best model has been chosen according to the highest C-score whose value of -0.62 suggests a correct topology. Likewise, the best model of Phyre2 has been obtained according to the identity percentage and confidence values which were 41% and 100%, respectively. This model was built with a template of non-phosphorylated STAT1 (PDB code: C1YVIB). Both the best model of I-TASSER and Phyre2 were compared by structural quality to distinguish the model to be used in next analyses. The results in the table 4.2 have shown that STAT2's model of I-TASSER obtains a quality factor of 70.74% better than Phyre2's model with a value of 56.65% according to Verify-3D. With ERRAT the trend is similar, because I-TASSER's model represents a better value than Phyre2's model, however this latter one has obtained an unacceptable value of 21.69% unlike to the value of 77.80% achieve by I-TASSER's model. For analysis with VADAR since that is a web server with different algorithms that performs a massive test of the three-dimensional model. It has been chosen values corresponding to α-helix, coils and β-sheet conformations that based on the three-dimensional model for both I-TASSER's and Phyre2's models have almost remained in the range of secondary structure gathered with the amino acid sequence (see table 4.2). Moreover, VADAR has permitted to calculate the free energy of folding which is the energy to keep folded or unfolded a protein, thus it has been observed that for energy of Phyre2's model is of -693.89kcal$\cdot$mol^{-1} which is lower than I-TASSER's model which is -671.47kcal$\cdot$.mol^{-1}. Both have been closed to the expected values of the software (-822.51kcal$\cdot$.mol^{-1} and -826.47kcal$\cdot$.mol^{-1}, respectively) (see table 4.2). On the other hand, Ramachandran plots were obtained by PROCHECK. Phyre2's model has reached more appropriated values for favored and allowed regions than I-TASSER's model since gathers values of 75% and 17.2% which are higher than the I-TASSER's model with values of 72.6% and 21.6%. However, Phyre2's model has gathered a higher percentage in the disallowed regions indicating that the stereo-chemical quality is inferior to I-TASSER'S model (see Fig. 4.3). Based on these outcomes were selected I-TASSER's model because their residues are located in the most favorable three-dimensional structure environment in agreement with the experimental functional domains [13], [25] as shown in the Fig. 4.5. In addition, there take into account the limitation and advantages of both approaches. Modeling by homology is usually more accuracy wherein stereo-chemical restrains and segments matching are then considered. However, it is limited to a template which must have a high identity with target sequence since with a percentage below 30% the accuracy of homology models decreases because of alignment error and incapability to generate structures that fit with the target sequences [26], [27]. Otherwise, modeling by threading is capable to build a model without a template which represents an improvement when the target sequence is unknown or has segments (surface loops) that are misaligned with a template sequences. Therefore, this latter is considered more successful in contrast to an approach by homology [26], [27].

The structures of each model improved through 10 ns of MD simulations (see table 4.2). However, Phyre2's model do not display a superior recovery since the trend is similar to the initial coordinates. All analyses above mentioned were also performed for NS5, although the three-dimensional model was obtained from protein data bank (Fig 4.5). According to Verify-3D and ERRAT, the quality factors of NS5 are higher than STAT2 model (see table 4.2). The analysis of secondary structure in NS5 by atomic coordinates have conformations more balanced in their composition, in contrast to the values obtained by amino acid sequence. The free energy of folding of their structure is $-782.28 \text{kcal} \cdot \text{mol}^{-1}$ which is still more stable than STAT2's models (see table 4.2). Ramachandran plot of NS5 displays that both the favored and allowed regions have fairly reached higher values getting a great quality model (see Fig 4.4). Likewise, NS5's model was subjected to a refinement with MD simulations which has considerably improved their structural quality in all parameters checked above.

Additionally, STAT2's model has been compared with experimental fragments of STAT2 in order to observe the structural similarity, as shown in the Fig 4.6. Each fragment (2KA4 and 5OEN) was aligned with STAT2's model where the fragment 2KA4 has been matched by 39 residues with a TM-score of 0.34 in the Transactivation domain while the fragment 5OEN has been paired by 168 residues with a TM-score of 0.86 in the Coiled-coil domain. Therefore, STAT2's model has hardly coincided with the fragment 2KA4 suggesting that this model is rather suitable to be used as a reference protein model of STAT2.

Table 4.2. Analysis of global quality of structure, secondary structure and free energy of folding to the initial and final coordinates with 10 ns of simulation of NS5 and STAT2 through Verify-3D, ERRAT, and VADAR.

Models	Verify-3D (%)	ERRAT (%)	VADAR			
			Free energy of folding (kcal·mol^{-1})	α-Helix (%)	β-Sheet (%)	Coil (%)
STAT2/I-TASSER's model	70.74	77.80	-671.47	41.00	15.00	42.00
STAT2/Phyre2's model	56.65	21.69	-693.89	38.00	13.00	48.00
NS5/ PDB code: 5TMH	82.61	93.61	-782.28	39.00	20.00	39.00
STAT2 with 10 ns	86.23	85.07	-	-	-	-
NS5 with 10 ns	92.75	84.68	-	-	-	-



Plot statistics

a

Residues in most favoured regions [A,B,L]	561	75.0%
Residues in additional allowed regions [a,b,l,p]	129	17.2%
Residues in generously allowed regions [~a,~b,~l,~p]	27	3.6%
Residues in disallowed regions	31	4.1%
	----	------
Number of non-glycine and non-proline residues	748	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	40	
Number of proline residues	57	

Total number of residues	847	

b

Residues in most favoured regions [A,B,L]	545	72.6%
Residues in additional allowed regions [a,b,l,p]	162	21.6%
Residues in generously allowed regions [~a,~b,~l,~p]	29	3.9%
Residues in disallowed regions	15	2.0%
	----	------
Number of non-glycine and non-proline residues	751	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	40	
Number of proline residues	58	

Total number of residues	851	

c

Residues in most favoured regions [A,B,L]	612	81.5%
Residues in additional allowed regions [a,b,l,p]	110	14.6%
Residues in generously allowed regions [~a,~b,~l,~p]	19	2.5%
Residues in disallowed regions	10	1.3%
	----	------
Number of non-glycine and non-proline residues	751	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	40	
Number of proline residues	58	

Total number of residues	851	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

Figure 4.3. Ramachandran plot of STAT2 model by ITASSER and Phyre2. (a) Phyre2's model (b) I-TASSER's model (c) STAT2's model with 10 ns MD simulation. The most favored regions are red. Allowed, generously allowed, and disallowed regions are yellow, pale yellow, and white, respectively. High stereo-chemical quality is considered when any residue lies in the disallowed region.

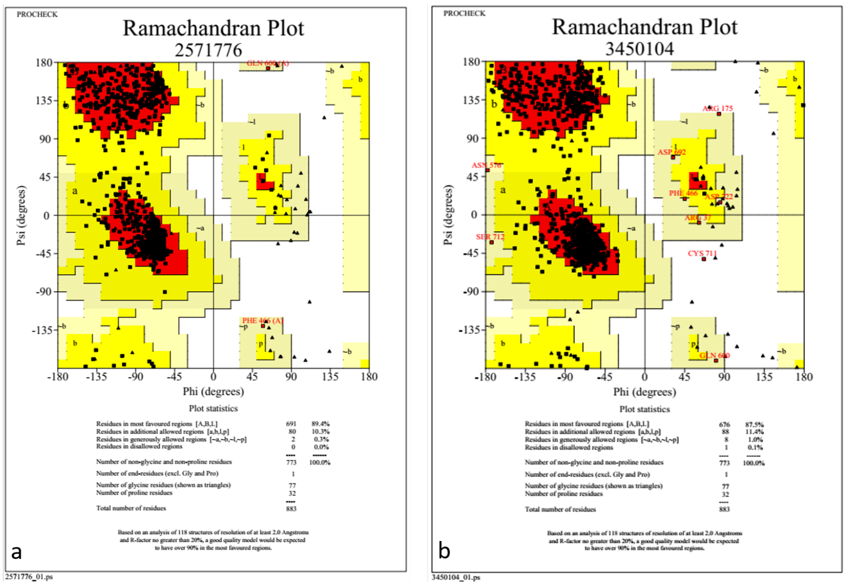


Figure 4.4. Ramachandran plot of NS5 model (PDB code: 5TMH). (a) NS5 without MD simulation (b) NS5 with 10 ns MD simulation. The most favored regions are red. Allowed, generously allowed, and disallowed regions are yellow, pale yellow, and white, respectively. High stereochemical quality is considered when any residue lies in the disallowed region.

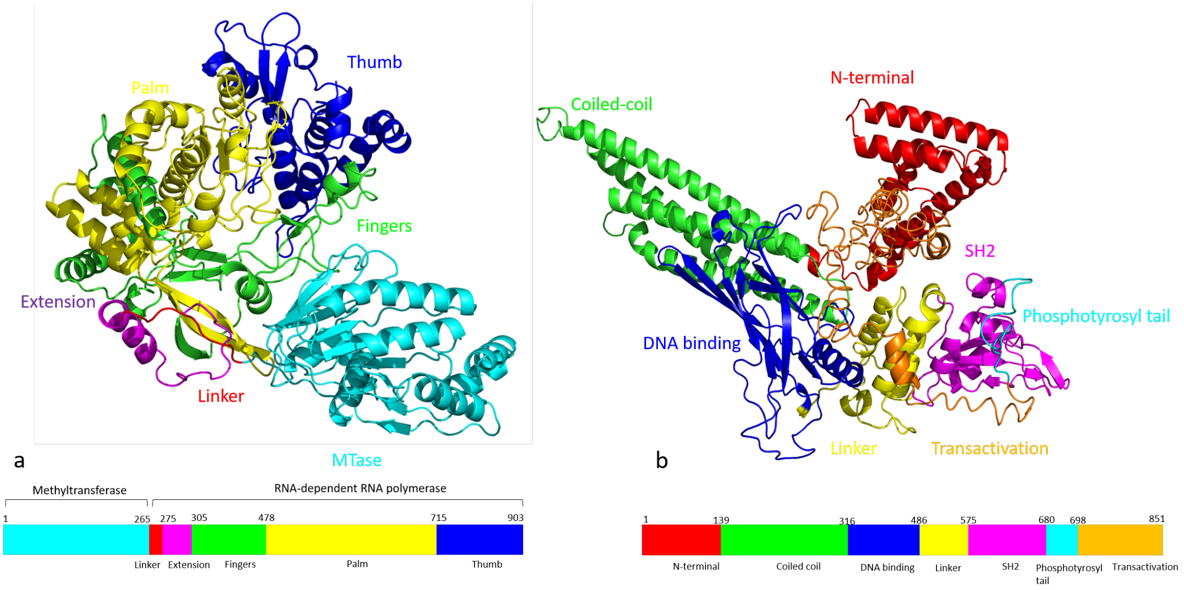


Figure 4.5. Three-dimensional structures of (a) NS5 and (b) STAT2 model. The functional domains of each model are shown in different color. NS5: Mtase = Methyltransferase domain (1-264 aa), Linker domain (265-275 aa), Extension (275-304 aa), Fingers domain (305-477 aa), Palm domain (478-714 aa) and Thumb domain (715-903 aa) STAT2: N-terminal domain (1-138 aa) involved in dimerization/-tetramerization, Coiled-coil domain (139-315 aa) involved in interaction with other proteins, DNA binding domain (316-485 aa), Linker domain (486-574 aa), SH2 domain (575-679 aa), Phosphotyrosyl tail segment (680-697 aa), and Transactivation domain (698-851 aa).

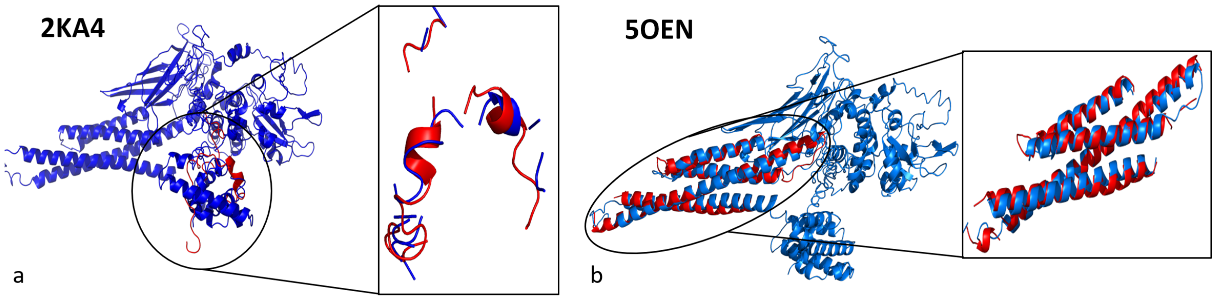


Figure 4.6. Structural comparison among the experimental fragments of STAT2 (PDB codes: 5OEN and 2KA4) with STAT2's model. (a) Fragment 2KA4 and STAT2's model (b) Fragment 5OEN and STAT2's model.

4.3 MD simulation of NS5 and STAT2

In order to understand the biological functions as well as interactions of these proteins, it is essential to explore the conformational changes on short time lapse at atomic level. MD simulation is a tool that achieves this aim since it provides a description of motion, structural properties and thermodynamic behavior of the systems (NS5 and STAT2) at equilibrium [29], [30], [33]. Therefore, NS5's x-ray structure and STAT2's model were subjected to 50 ns MD simulation protocol defined in the method section to minimize and equilibrate at physiological conditions. Several system attributes were calculated of the NS5 and STAT2 trajectories. RMSD has been monitored to observe the equilibrium of systems where NS5 ($\overline{x}$=0.25 nm ± 0.023) has rapidly achieved a stable steady state, however STAT2 ($\overline{x}$=0.83 nm ± 0.089) has reached it after 5 ns, as shown in the Fig. 4.7. Moreover, NS5 has a low displacement of its atoms during the simulation which have kept in a range of 0.2 to 0.3 nm. In contrast, the displacement of STAT2 has been closed to 1 nm with a range of 0.8 to 1 nm during the period of stabilization. Disordered regions can be involved in the inability of STAT2 to reach a rapid stable state since the number of them is higher than NS5. Moreover, the internal motions in the protein may gather many conformations on these regions by their known flexibility. Likewise large atomic displacements are achieved by surfaces residues which correspond with structural details of STAT2 obtained previously in the sequence analysis [33].

Root Mean Square Fluctuation (RMSF) of Cα atoms for both NS5 and STAT2 have been analyzed in the Fig. 4.8. NS5 ($\overline{x}$=0.16 nm ± 0.063) is more stable since that has a low fluctuation of residues in a range of 0.1 to 0.5 nm, although the region between 1-250 residues may be considered as a zone with high fluctuation in contrast to the rest of residues. It has been shown a correlation with the disordered residues predicted by sequence analysis. Otherwise, STAT2 ($\overline{x}$=0.30 nm ± 0.123) has shown an elevated atomic fluctuation in the initial and terminal region of the protein which has displayed certain correspondence with the sequence-based prediction analysis where the terminal region specially had the highest fluctuation and disorder. Therefore, these regions are conformationally more flexible compared to remaining residues and also may be involved in complex formation with other proteins, substrate/inhibitor binding which are relevant in the viral response [76]. Furthermore, the high fluctuation in these regions are associated with side chains which have anharmonic motions at 300 K and spend much of their time in the minimum energy position before to hop a barrier of transition to another minimum. In contrast to residues of NS5 which executes small fluctuations because they hop a barrier of transition quite rapid [33]. With respect to the structural behavior both NS5 and STAT2, the α-helix and loop displacements lead to structural differences of the backbone. At the same time, the motion of side chains produce reorientations that together to coiled coils (loops) rearrangements generate α-helix displacements as it is observed in the initial region of STAT2 included to the N-terminal and part of Coiled coil domains. Likewise, the side chain transitions and dihedral angle transitions in the loop regions produce loop motions that correspond to the high fluctuation of Phosphotyrosyl tail and Transactivation domians in the last region of STAT2 [33].

Radius of gyration (Rg) has been used to measure the compactness of the protein structures of NS5 and STAT2 (see Fig. 4.9). During the simulation, NS5 and STAT2 have displayed a steady state, but has been noticed that STAT2 ($\overline{x}$=3.61 nm ± 0.023) is more extended protein, contrary to NS5 ($\overline{x}$=3.24 nm ± 0.025) which is slightly more compressed. This behavior is related with backbone and side chains in the protein since the motion of side chains plus the rearrangements of loops produce α-helix fluctuations, hence the compactness of STAT2 is reduced by the high presence of α-helix. In NS5, the compactness is high because there is great range of α-helix packing by changes in the side chain volumes suggested in globular structures [33].

Solvent Accessible Surface (SAS) has been measured for each protein. SAS of NS5 ($\overline{x}$=400 nm^2 ± 3.90) is stable during the simulation in a range of 390 nm^2 to 410 nm^2. While SAS of STAT2 ($\overline{x}$=465.45 nm^2 ± 15.67) is initially higher but during the simulation has changed since that it has descended from roughly 480 nm^2 until 440 nm^2 (see Fig. 4.9). These latter results agree with the pioneer shape idea of proteins since STAT2 had a fibrous form with more access to solvent unlikely to NS5 which is a globular protein and the access to the solvent is much more restricted.

From the trajectories of NS5 and STAT2, it was performed a clustering analysis to explore the conformation of proteins generated by MD simulation. The GROMOS clustering algorithm with a RMSD of Cα cut-off was employed in order to determine the structurally similar cluster and obtain a representative structure of both proteins. According to the RMSD cut-off detailed in the methods section, the dominant clusters for NS5 and STAT2 constitute ~75% of total protein structures. Therefore, it was extracted representative structures from these cluster that will be employed in the following analyses since they have the lowest stabilizing energy which is considered that these structures are closed to a native structure [25], [83], [84], [85].

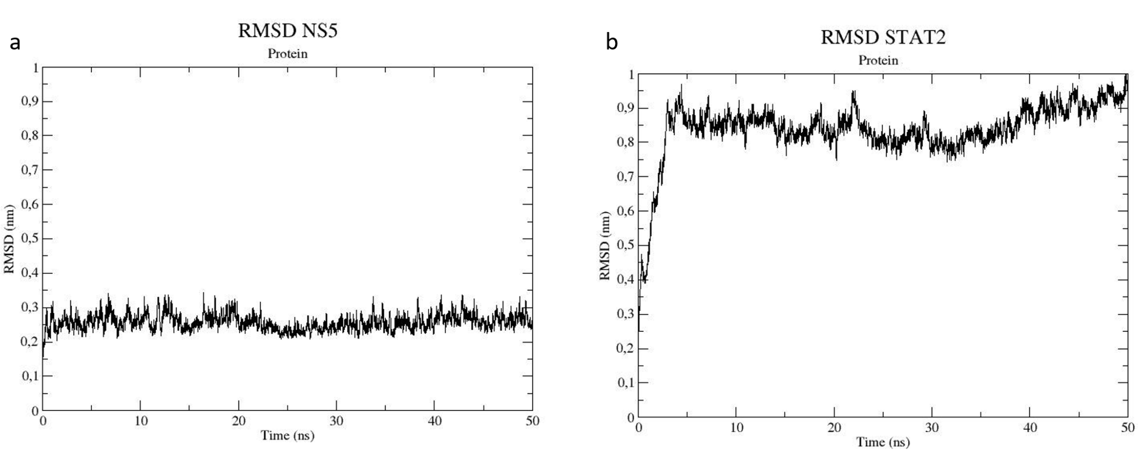


Figure 4.7.RMSD for NS5 and STAT2 as a function of time. (a) RMSD evolution of NS5 (b) RMSD evolution of STAT2.

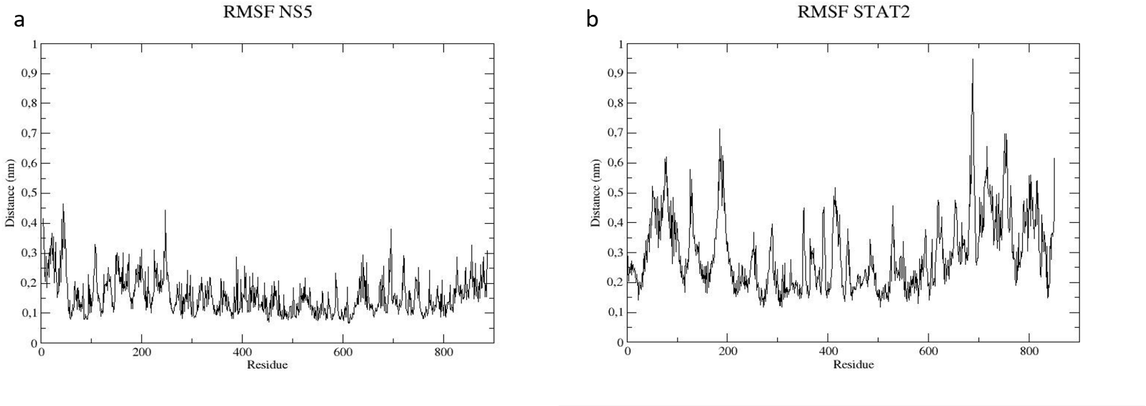


Figure 4.8.RMSF of the backbone Cα atoms versus the number of residues present in each sequence both (a) NS5 as (b) STAT2.

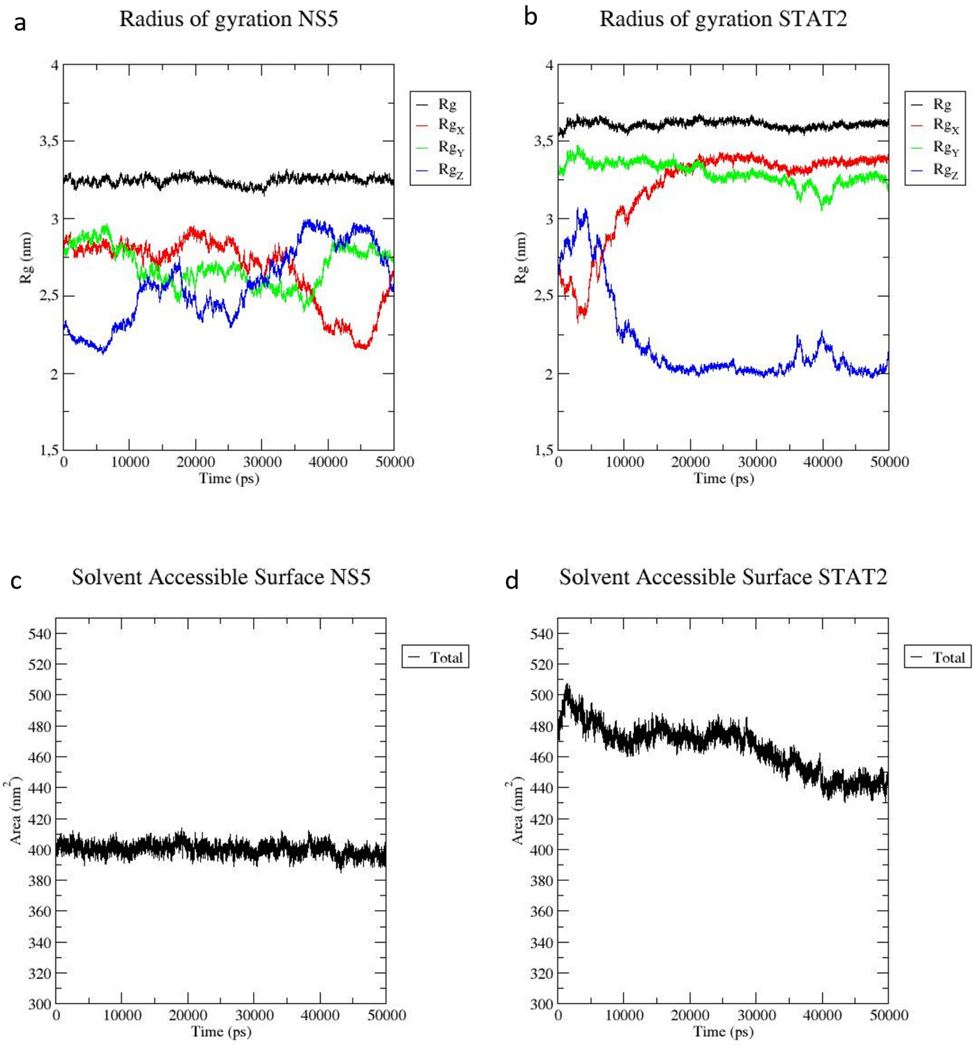


Figure 4.9. Radius of gyration (Rg) and solvent Accessible Surface (SAS) for NS5 and STAT2 as function of time. (a) Rg evolution of NS5 (b) Rg evolution of STAT2 (c) SAS evolution of NS5 (d) SAS evolution of STAT2 during 50 ns.

4.4 Analysis of docking complex NS5-STAT2

The previous MD simulation of NS5 and STAT2 is employed to improve the virtual screening in the docking processes since it can be considered as step of refining such models. This refining provides flexibility and a structural rearrangements to proteins since the methods of protein-protein docking are usually based on the shape complementarity. Therefore, MD simulation allows to obtain a more realistic structure of each model [28], [41], [42], [43]. From trajectories of MD simulations has been extracted a representative structure of NS5 and STAT2 through clustering. These structures are used as target to build a ensemble docking [28].

Experimental studies have suggested that the interaction NS5-STAT2 leads the degradation of STAT2 by NS5, but the specific sites of interaction are still undefined [8], [10], [11]. Therefore, molecular docking permits to study the roles and functions in interaction's ligand-receptor [41]. In order to explore the interaction between these proteins, docking analyses have been performed through ClusPro and Pydock. ClusPro provides a list of clusters with their lowest energy and representative members. In order to select the ClusPro's model, ClusPro provides four scoring functions, one of them is called Balance which is suggested by the program itself when there is not information about of interaction in the interface of proteins. From the 29 clusters, the first ten were selected based on the largest number of members because of the lowest energy parameter is not an indicator to select the best model of docking [64]. The models obtained by ClusPro are shown in the Fig 4.10. Regarding to Pydock, the ten first models have been selected according to the total stabilizing energy proposed by the web server itself, as shown in the Fig. 4.10. At this stage, two troubles arises, the first is scoring function of ClusPro and Pydock and second is the binding site which is unknown to predict the correct solution [41]. Although some algorithms are able to score a list of possible structures, however they are not reliable to discriminate false positives, that is, complexes with a high score but with a low rank [41].

Therefore, in order to classify of docked complexes were employed the software PPCheck with a process of discrimination by functional domains, and ΔG_{Bind}. PPCheck quantifies the strength of protein-protein interaction using pseudo-energies. As a result, the models of ClusPro and Pydock have been classified. The outcomes have shown that the total stabilizing energy for Pydock's models are positive ($\Delta E>0$) while that for ClusPro's models are negative ($\Delta E<0$). The total stabilizing energy values that have negative energy tendency are related with the increase of number of interface residues [66]. Hence, a negative energy suggests that there is contact interface between proteins. In contrast, the interaction is deficient due to the contact areas are almost null in systems with positive energy (see table 4.3). Among of them, the ClusPro's models 4, 7 and 9 have obtained the largest total stabilizing energy (see Fig. 4.11). One way to be precise in the prediction of docking pose and to know whether docked complex has a stable interface is through the interface residues which should be within of the range of 51-150 and a normalized energy per residue from -2kJ$\cdot$mol^{-1} to -6kJ$\cdot$mol^{-1} [66]. In our models selected with the largest total stabilizing energy, the normalized energy per residue has also been closed to -6kJ$\cdot$mol^{-1}, therefore they have fallen in a correct docking pose (see table 4.3). ClusPro's models have shown to interact via known domains in both proteins. A test of total stabilizing energy and normalized energy per residue among the N-terminal domain from STAT2 and Mtase and Thumb domains from NS5 have been performed. The models 1, 4, and 7 have demonstrated that the interaction among these domains have gathered the largest total stabilizing energy, however the interaction in the model 9 is only between the N-terminal and Mtase domains (see table 4.4). Likewise, the models 1, 4 and 7 have reached the largest normalized energy per residue close to -6kJ$\cdot$mol^{-1} showing that these models involve a correct docked position. In contrast to the other models that do not show interaction among these domains. In the interest of contrasting the total stabilizing energies obtained by each model and their domains, the Protein-Protein Interactions in Macromolecular Assemblies (PIMA) tool was used [86]. This one has a database called PIMADb [87] of 60.555 entries of protein-protein interactions of protein assemblies with which our results have been compared. The values corresponding to the total stabilizing energy of the interaction of N-terminal/Mtase and N-terminal/Thumb are along of the known interactions from other complexes (see Fig. 4.12).

In order to understand the difference between the model of ClusPro and Pydock, variables as the surface area provide features in the protein-protein interaction which displays a high degree of structural complementarity and chemical complementarities [88]. The PISA web server was then employed in which is possible to calculate the interface area (Å^2) and the solvation free energy (ΔG_{Solv}), displayed in the Fig. 4.13. As a result, the interface areas of ClusPro's models ($\overline{x} = 1973.59\text{Å}^2$) are larger than any Pydock's model ($\overline{x} = 1068.03\text{Å}^2$). Similar results to ClusPro's models have been found in a set of 75 crystal structures with an interface area average of 2000Å^2 in each member which is considered a specific protein-protein interaction with high complementarity [88], [89]. It has also been observed that protein-protein interaction has large contact surfaces (1500-3000 Å^2) while that the contact area between small molecules and proteins targets has been estimated between 300 to 1000 Å^2 [90]. These outcomes are connected with solvation free energy ($\Delta G_{Solv}<0$) which are higher in most of ClusPro's models with models 8 and 9 being an exception (see Fig. 4.13). Therefore, the interaction between protein-protein lead an increment of interacting area and solvatation free energy (ΔG_{Solv}), because of these values are associated to a better affinity between proteins. However, this is done without taking into account the effect of hydrogen bonds and salt bridges across the interface area which has consequently been considered as partial outcomes to discriminate between ClusPro and Pydock models.

The Gibbs binding free energy (ΔG_{Bind}) has been calculated using empirical force field of FoldX for atomic coordinate of ClusPro's models. In this initial analysis each model has achieved a value of positive of interaction energy $\Delta G_{Bind}>0$, as shown in the Fig. 4.14. It means that the docked complexes may be considered as unstable structures since it denotes energy-unfavorable coupling between both proteins. This unstable state between proteins NS5-STAT2 can also be considered as an non-covalent interaction with negative cooperativity since the affinities between ligand-receptor are decreased, that is, the docked complex is less well bonded and exhibits their atoms major internal motion [91]. A negative cooperativity suggests that the docked complexes will need a great amount of energy to coupling because the binding between NS5 and STAT2 is not spontaneous, that is, a process highly endothermic. Moreover, ΔG_{Bind} has two contributions enthalpic and entropic wherein a $\Delta G_{Bind}>0$ is related with a favorable change entropy [91]. Additionally, the non-covalent interaction with negative cooperativity in the interface between NS5-STAT2 may be influenced by coupling sites geometry since the contact distance between ligand-receptor is greater than a coupling of non-covalent interaction with positive cooperative causing a structural loosening in the docked complex [91]. The models 1, 7, 4, and 9 show the largest binding energy (ΔG_{Bind}) in comparison with the remaining models. However, it has only been decided to take into consideration models 1, 4, and 7 to be subjected to MD simulation because they showed interaction in the domains such as Mtase and Thumb with N-terminal in the NS5-STAT2 complex.

Table 4.3. Results of total stabilizing energy and normalized energy per residue for the ten models of ClusPro and Pydock computed by FoldX.

	Total stabilizing energy (kcal·mol^{-1})		**Normalized energy per residue (kJ·mol^{-1})**	
Models	**ClusPro**	**Pydock**	**ClusPro**	**Pydock**
model 1	-77.942	16.066	-2.22	0.35
model 2	-99.993	153.437	-2.25	4.59
model 3	-80.707	167.854	-2.24	4.65
model 4	-139.469	109.496	-3.15	4.16
model 5	-79.381	134.481	-2.27	5.52
model 6	-89.536	135.026	-2.18	5.88
model 7	-121.016	-9.668	-2.84	-0.4
model 8	-60.770	126.955	-1.72	4.05
model 9	-125.660	43.442	-3.23	1
model 10	-88.556	25.808	-1.87	0.92

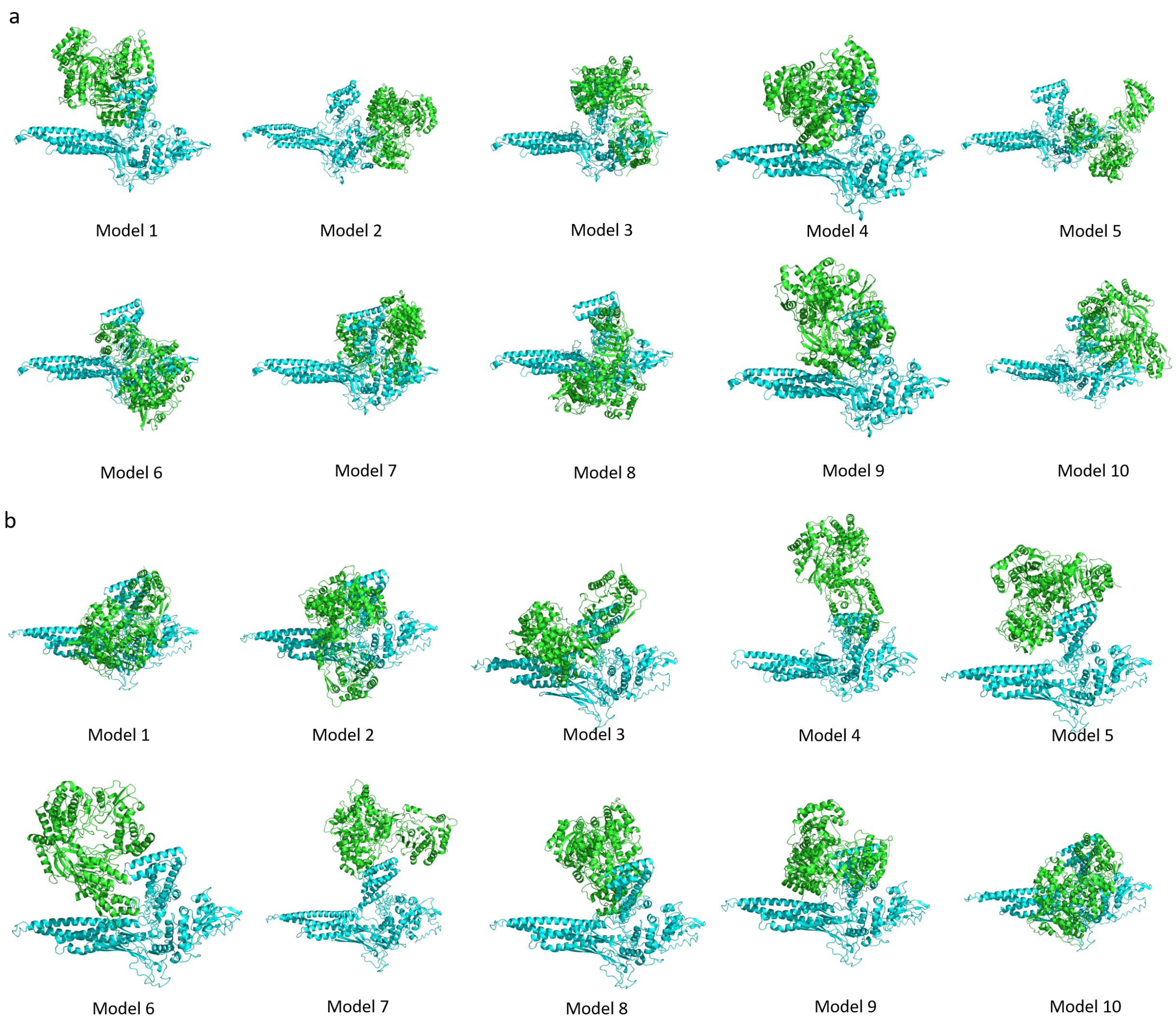


Figure 4.10. The first ten models of NS5-STAT2 complex obtained by (a) ClusPro and (b) Pydock.

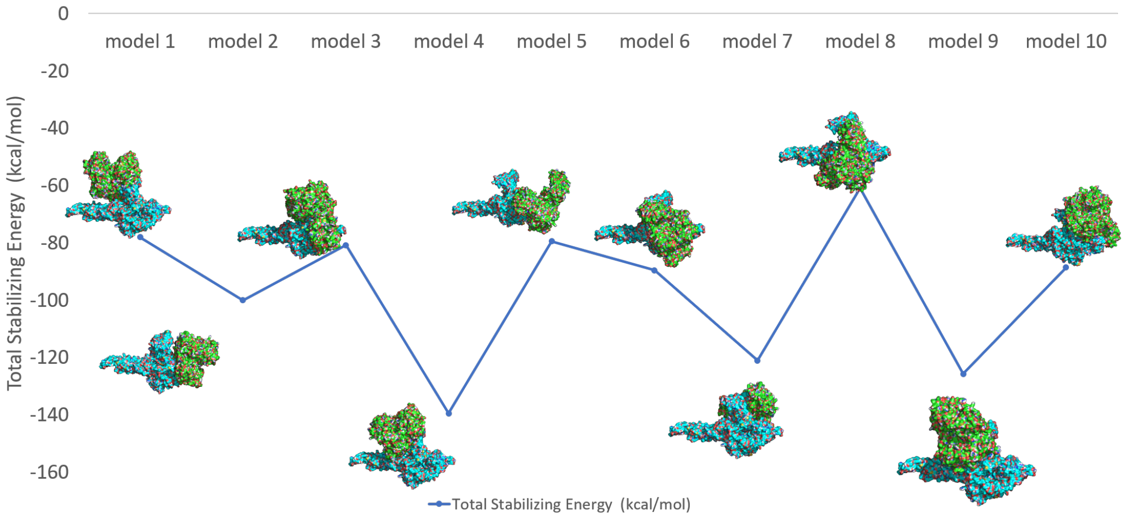


Figure 4.11. Results of PPCheck web server for total stabilizing energy in ClusPro's models.

Table 4.4. Results of total stabilizing energy and normalized energy per residue for ten ClusPro's models based on the functional domains of NS5 and STAT2. The functional domains involved in the interaction of complex were the N-terminal from STAT2 and Mtase and Thumb from NS5.

	Total stabilizing energy (kcal·mol^{-1})		Normalized energy per residue (kJ·mol^{-1})	
Models	**N-Ter/Mtase**	**N-Ter/Thumb**	**N-Ter/Mtase**	**N-Ter/Thumb**
model 1	-23.2	-27.192	-2.7	-1.9
model 2	-	-	-	-
model 3	-	-	-	-
model 4	-19.116	-48.282	-3.33	-3.61
model 5	-	-	-	-
model 6	-9.037	-	-1.18	-
model 7	-59.388	-39.959	-3.5	-3.89
model 8	-3.774	-	-0.88	-
model 9	-1.114	-	-4.66	-
model 10	-	-	-	-

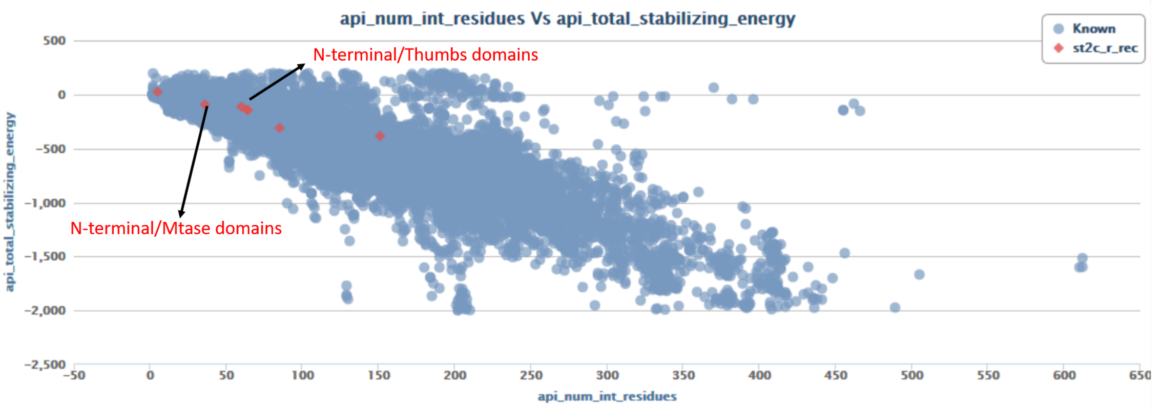


Figure 4.12. Plot of total stabilizing energy from the query domains (marked with red) with the known interactions (marked with blue) from other complex obtained from PIMA.

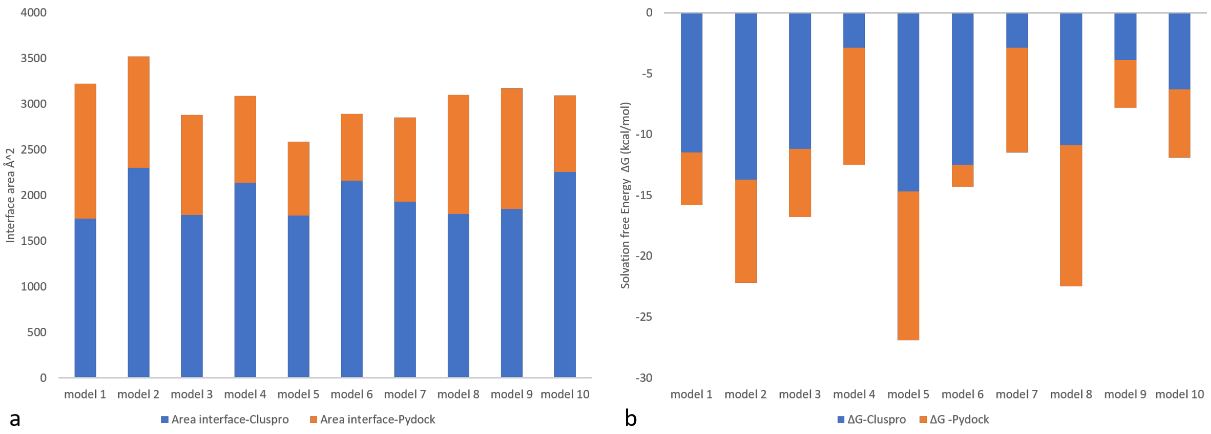


Figure 4.13. Results obtained through PISA web server. (a) Interface area (Å^2) and (b) solvation free energy achieved by the formation of the interface for ClusPro and Pydock.

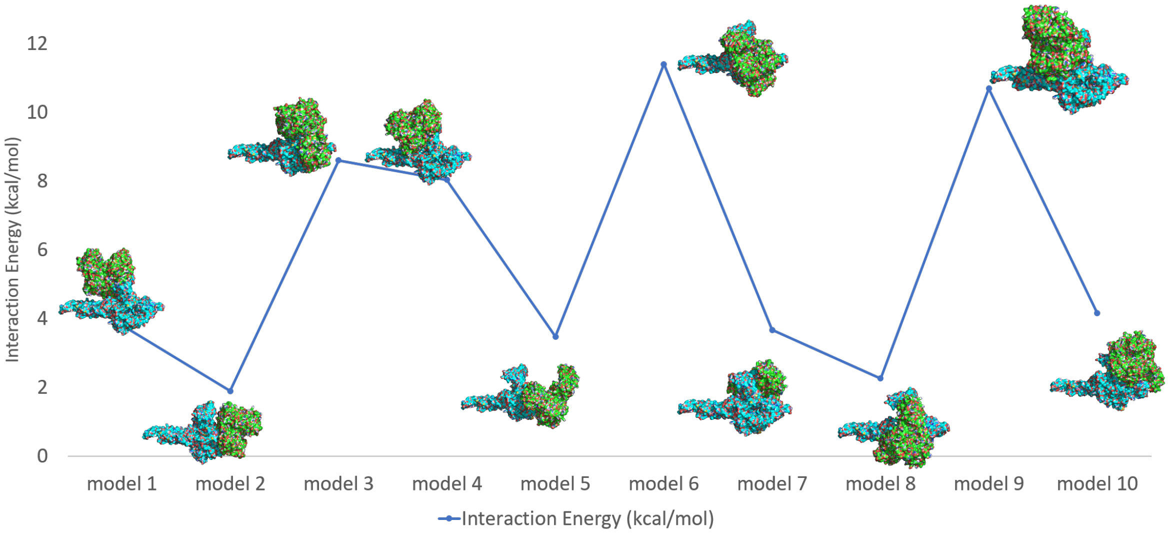


Figure 4.14. Results of Gibbs free energies of binding for ten ClusPro's models calculated by FoldX software.

4.5 Contact map and electrostatic analysis

In order to confirm the interaction among Mtase and Thumb with N-terminal domains, a contact map of intramolecular interactions of the NS5-STAT2 complex is illustrated in the Fig. 4.15. The distance range between two proteins have been marked with red, yellow, green and blue for 7Å, 10Å, 13Å, and 16Å, respectively. Largest regions of two partners that are in contact have been located in the map [70]. The regions associated corresponds to first residues of Mtase domain (1-200) and the last residues of Thumb domain (650-750) in the NS5 while that for region of STAT2 is in the residues of N-terminal domain (1-150) and a smaller interaction region in the residues from 300 to 400 that correspond to DNA binding domain. The contact map confirms our previous results since the protein-protein interaction is mainly focused on in the domains described.

Additionally, an electrostatic potential analysis has been performed because electrostatic interactions are favored for the protein-protein interaction as well as the stabilization in complex. The electrostatic role depends on the type of hetero- or homo- complexes, that is, a complex is formed by different or identical proteins which have net charge (positive or negative). In our case, the complex NS5-STAT2 is a heterocomplex which carries an opposite net charge that lead to attraction between proteins. However, the arrangement of heterocomplex will be limited by residues distribution with their charges that will change global net charge of each protein when the residues will be at short distances in the interface [92]. Fig. 4.16 shows the NS5-STAT2 complex protein. In the region of N-terminal domain of STAT2, their buried residues are highly polar. In contrast, the residues located in the cavity of NS5 have a hydrophobic character with few polar groups around. Studies suggest that polar residues in the interface favorably contributes in two ways the first through an specific association between proteins and second improvement the stabilization of complexes since the interacting forces do not need to be strong for the formation of complexes [93]. Moreover, regions in the protein with polar groups are usually sites known as hot spots which are crucial for a better affinity between proteins [92], [93], [94].

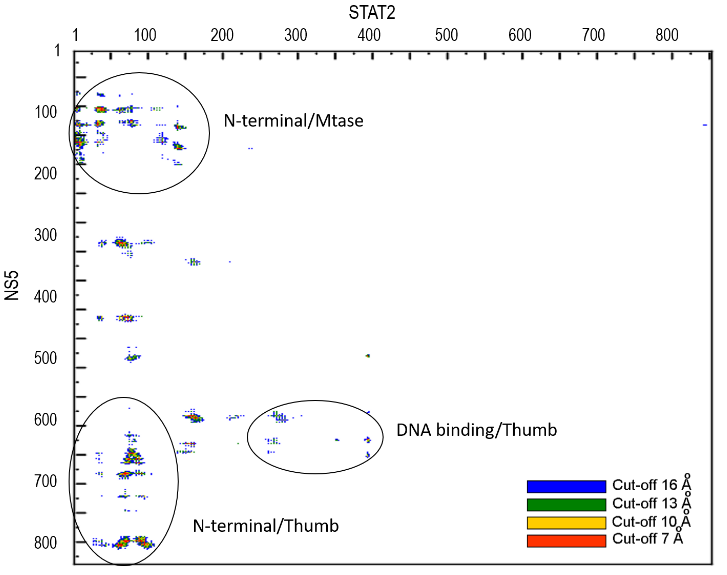


Figure 4.15. Contact map of NS5 and STAT2 shows the intermolecular contacts at reducing distances which are red=7Å, yellow=10Å, green=13Åand blue=16Å. The circles identify the domains related such as N-terminal/Mtase, N-terminal/Thumb and Thumb/DNA binding.

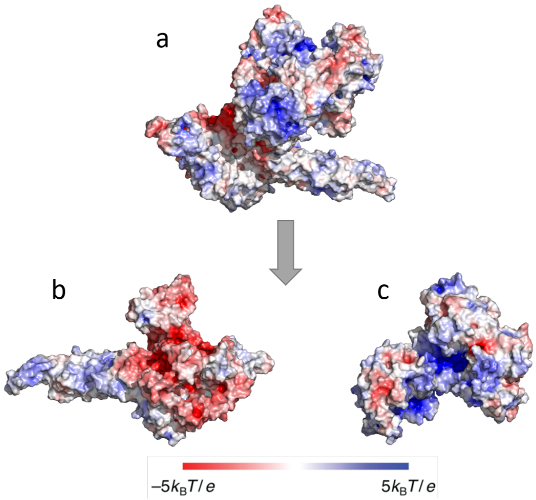


Figure 4.16.A view of electrostatic surface potential for (a) NS5-STAT2 complex, (b) STAT2 and (c) NS5. Red is negative charge, blue is positive charge, and white is neutral.

4.6 MD simulation of NS5-STAT2 complexes

The models 1, 4, and 7 have been minimized in order to optimize the complex geometry and check the binding energy (ΔG_{Bind}) again. The complexes were analyzed for their structural integrity by RMSD, RMSF, Rg and SAS. RMSD has been checked for each protein forming the complex (see Fig 4.17). NS5, for all complexes, have shown that its stability has slightly been affected during the MD simulation, although the equilibrium has rapidly been reached within 50 ns. The models have displayed a similar range to the NS5 unbounded, but it is noteworthy that NS5-model 4 ($\overline{x}$=0.28 nm ± 0.038) from 35 ns has suffered a change in its equilibrium. In contrast, the stable state has been difficult to achieve for STAT2 in all complexes. STAT2 bounded has showed a RMSD lower than STAT2 unbounded. For STAT2-model 1 ($\overline{x}$=0.49 nm ± 0.143) just after 40 ns, an equilibrated state has barely achieved, although it is necessary more time of simulation. STAT2-model 4 ($\overline{x}$=0.57 nm ± 0.119) has a great variation before 20 ns, however after this time has reached an equilibrium state. STAT2-model 7 ($\overline{x}$=0.48 nm ± 0.081) is quite stable during the simulation. The behavior of STAT2 is also more unstable due to the disordered regions that have an great influence in the complex formation. RMSF for the NS5 and STAT2 bounded forms have shown that the atomic fluctuation has decreased during the MD simulation in comparison with NS5 and STAT2 unbounded forms. Nevertheless, STAT2 has presented a large fluctuation in the initial and final residues which are associated with the N-terminal and Transactivation domains, respectively. In the initial region corresponding to N-terminal domain has displayed a large atomic fluctuation which is conditioned by the interaction with NS5 since in the bounded forms their atomic fluctuation is reduced. In contrast, it is remarkable the high fluctuation in the residues corresponding to the Transactivation domain since it has direct relation with the disorder in this protein and its secondary structure which is integrate by coiled-coil in the all models (see Fig 4.18). Rg property was measured for the bounded forms of NS5 and STAT2 where the compactness of proteins has been affected since the interaction between them (see Fig 4.19). Particularly, the NS5-model 1 ($\overline{x}$=3.13 nm ± 0.012) shows a different behavior with respect to others since has an remarkable inferior Rg which means that the relaxation is conditioned by the interaction with STAT2. The same behavior of Rg shows for STAT2-model 1 ($\overline{x}$=3.25 nm ± 0.059) since the compactness is higher than the other models. In the case of NS5-model 4 ($\overline{x}$=3.21 nm ± 0.011) and NS5-model 7 ($\overline{x}$=3.24 nm ± 0.015) their trends are similar to NS5 unbounded since their values are within the range. For STAT-model 4 ($\overline{x}$=3.55 nm ± 0.026) and STAT2-model 7 ($\overline{x}$=3.53 nm ± 0.020) their trends are similar, despite their compactness is slightly affected by the interaction with NS5. As a consequence, the performance of STAT2 forming the complex has a higher range of gyration suggesting a less tight packing as compared to NS5. The last property observed has been SAS, where NS5-model 1 ($\overline{x}$=338.08 nm^2 ± 3.09) and STAT2-model 2 ($\overline{x}$=349.95 nm^2 ± 4.16) have a significant decrease in the access to solvent in comparison to the remaining models and unbound forms (see Fig 4.20). Hence, for model 1 may be favorable the interaction because of the distance of coupling between proteins is shorter suggesting that there is a greater rearrangement in the contact surface reducing the access to water molecules [95].

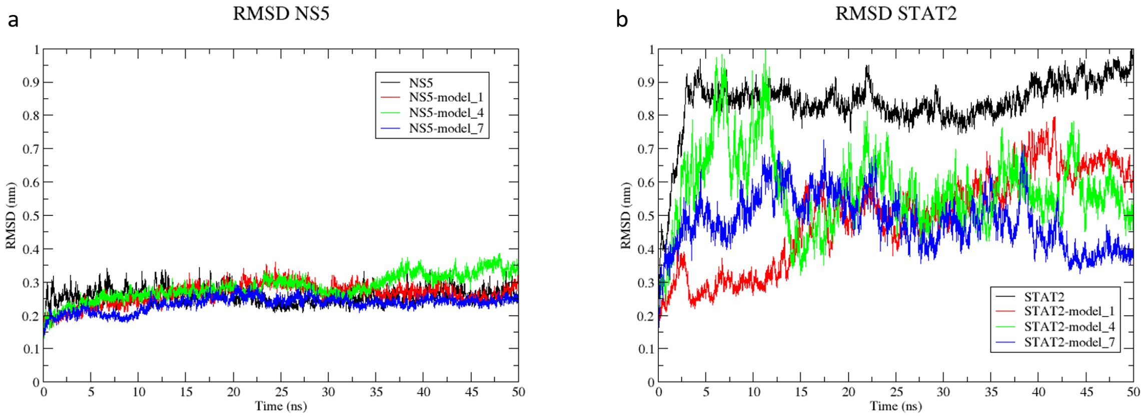


Figure 4.17.RMSD of bound forms of NS5 and STAT2 compared with the unbound form of each protein. (a) RMSD of NS5 (b) RMSD of STAT2 as unbound and bound forms

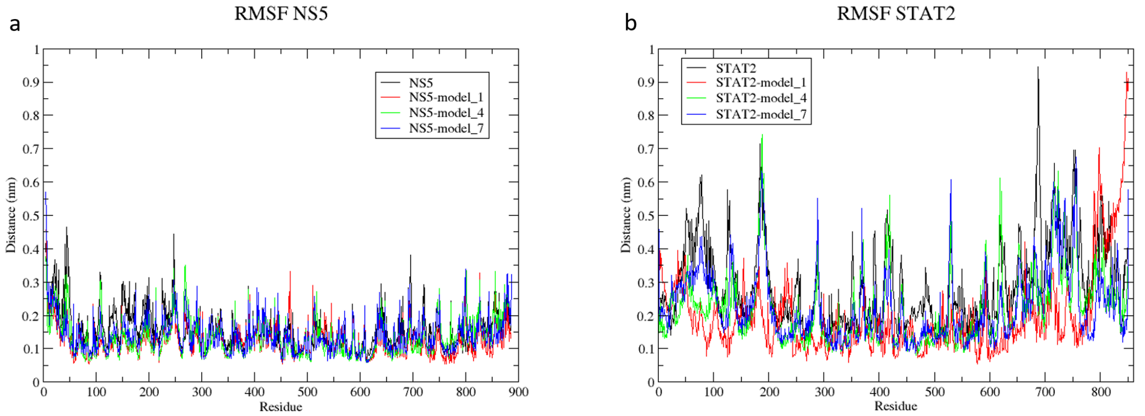


Figure 4.18.RMSF backbone Cα atoms of bound forms of NS5 and STAT2 compared with the unbound form of each protein. (a) RMSF of NS5 (b) RMSF of STAT2 as unbound and bound forms

In order to verify the behavior of the complex NS5-STAT2, the Gibbs binding free energy (ΔG_{Bind}) has been calculated for representative conformations (snapshots) of each trajectory in the models 1, 4, and 7. According to the clustering, each model has obtained 6001 structures which have been clustered in 33, 22, and 21 groups for models 1, 4, and 7, respectively. Through FoldX, the Gibbs binding free energy (ΔG_{Bind}) of each group was calculated.

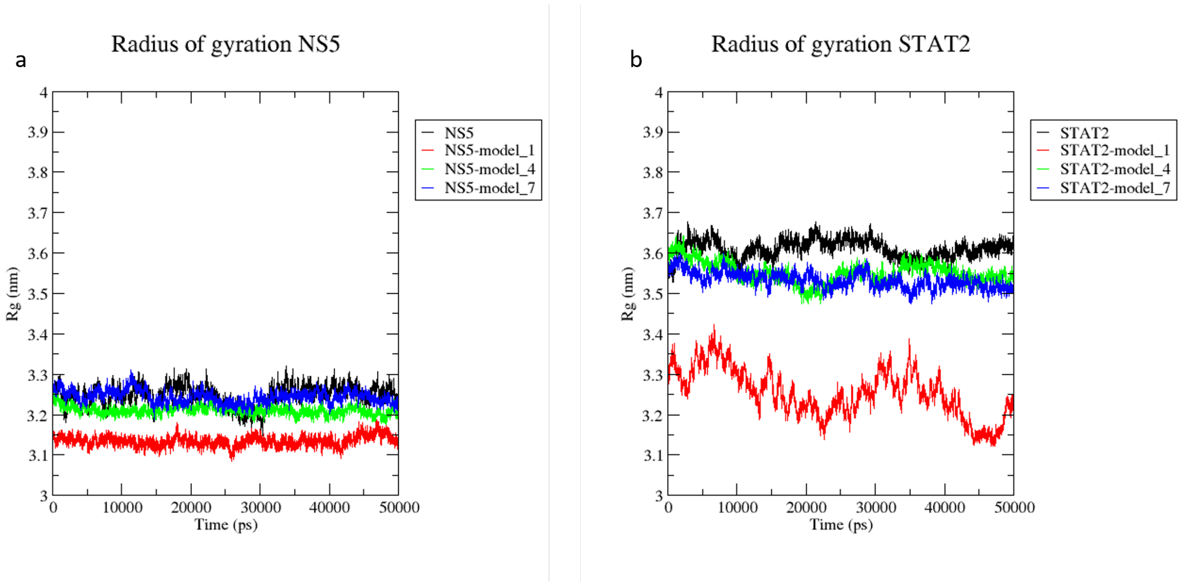


Figure 4.19.RG of bound form of NS5 and STAT2 compared with the unbound form of each protein. (a) RG of NS5 (b) RG of STAT2 as unbound and bound forms

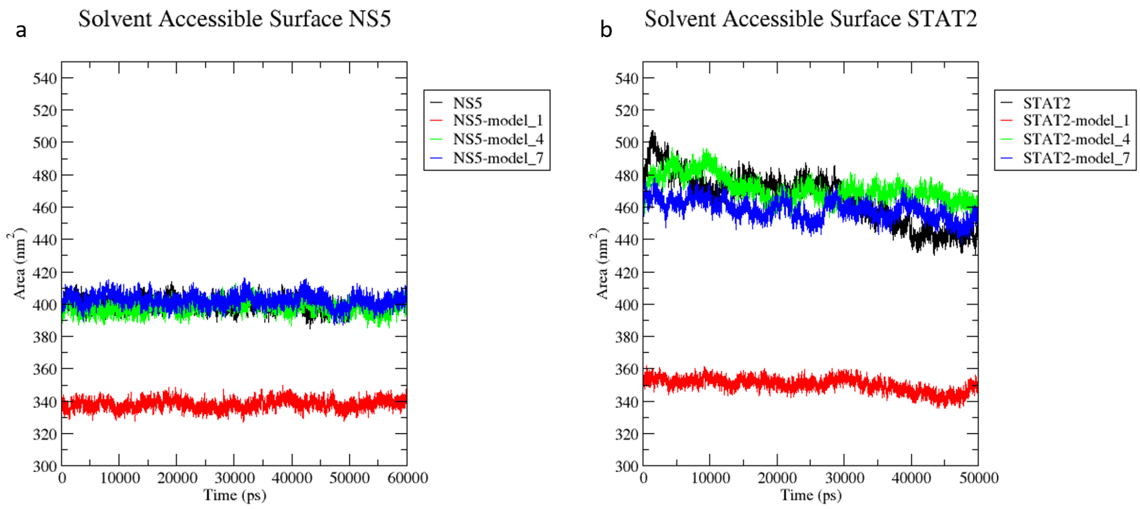


Figure 4.20.SAS of bound form of NS5 and STAT2 compared with the unbound form of each protein. (a) SAS of NS5 (b) SAS of STAT2 as unbound and bound forms

Hence, the results show that ΔG_{Bind} of model 1 and 7 (ΔG_{Bind}<0) are -4.30 kcal·mol^{-1} and -1.67 kcal·mol^{-1}, respectively while for model 4 (ΔG_{Bind}>0) is 0.27 kcal·mol^{-1}. The coupling between the proteins is energetically favorable in models 1 and 7 in opposition to model 4. The values of Gibbs binding free energy ΔG_{Bind}<0 are also obtained in the range of experimental measurements of binding affinity (kcal·mol^{-1}) of protein-protein tested on a bechmark of 144 complexes [96]. Likewise, a study with empirical approach based on three variables of the interface in complexes has estimated Gibbs binding free energy that are almost closed to the value of model 1 [97]. These latter results have demonstrated that by applying MD simulation to the system is possible to improve the values with respect the Gibbs binding free energy. The initial atomic coordinates of models have shown to reach a $\Delta G_{Bind}>0$ which mean that docked complex has non-covalent interactions with negative cooperativity, however after MD simulation, the models 1 and 7 have gathered a $\Delta G_{Bind}<0$ which benefit the non-covalent interactions with positive cooperativity [91]. This change in the cooperativity is acceptable in a same system since the motion of atoms in the MD simulation produces a new stable state at protein that has a net effect in the thermodynamic parameters [33], [91]. Hence, non-covalent interaction with a positive cooperativity is related with an exothermic binding which allows an increment in the bonding ligand-receptor. Likewise, a positive cooperativity is associated with a favorable enthalpy and adverse in entropy because with a strong coupling is reduced the internal motions of the complex improvement the non-covalent bonding [91]. In pathway of signal transduction factors exists proteins agonists that in some cases are beneficial in the bonding with other entities, because they have an adverse contribution of entropy (positive cooperativity), however they also may induce the dissociation with a receptor being adverse for coupling since they have a contribution favorable in entropy (negative cooperativity). Antagonists proteins as NS5 are species that bind to receptor as STAT2 but they do not active it. Thus, an antagonist protein is adverse to coupling since their contribution is favorable in entropy providing a non-covalent interaction with negative cooperativity [91]. These latter insights may explain our results since the initial atomic coordinates of models 1 and 4 displayed unfavorable non-covalent interaction, but after of MD simulation of these models show improvements Gibbs free binding energy that are favorable to the coupling. However, the values of ΔG_{Bind} associated to the models are still less in comparison with other studies [91].

4.7 Interaction NS5-STAT2

Viral infections are mediated by several protein-protein interactions where proteins are considered as nodes and their interactions as edges represented as networks [98]. Likewise, the protein-protein interaction is mediated by domain-domain interactions as has been understood in our analyses where it has been identified the interaction is given by N-terminal domain from STAT2 and Mtase and Thumb domains from NS5 [98]. Binding sites in the interface of NS5-STAT2 complex have a direct physical contact specially when there is an interaction non-covalent with positive cooperativity since the distance between ligand-receptor is shorter. Therefore, interaction between proteins have been checked through an analysis of residues in the protein-protein interface. The interfaces can be classified as endogenous or exogenous interface in the virus-host [98]. Particularly, the interaction NS5-STAT2 involves an exogenous interface since it is mediated between proteins belonging to distinct proteomes. Tools such as FoldX, PPCheck and PIC were employed to analyze the docked complex. FoldX also describes interface residues on the complex which are necessary to understand its protein-protein interaction since it may be used in the development of new antiviral therapies [98]. Interface between proteins interact through physical processes known as van der Waals interactions, hydrogen bonding, electrostatic interactions, hydrophobic interaction, exclusion of solvent, salt bridges, ionic interaction and entropic changes [96]. Residues related with interface in NS5-STAT2 have also been characterized through PPCheck and PIC.

A consensus of residues in the interface provided of ten ClusPro's model by FoldX has been performed, hence an analysis of classification has been performed to determine the frequency of residues that present any type of interaction in the protein interface of docked complex. It has been found that interactions in the interface of NS5-STAT2 are stabilized by electrostatic interaction, hydrophobic interaction, salt bridges and ionic interaction. In the case of NS5, 33 residues are in the contact area with an elevated frequency in all models. Moreover, 16 of them have presented one o more type of interaction described above. The residues involved are Arg-163, Arg-175, Arg-37, Arg-57, Arg-681, Arg-84, Arg-856, Glu-149, Leu-847, Lys-105, Lys-331, Pro-108, Pro-857, Trp-848, Val-335, and Val-336. These residues have located in the Mtase and Thumb domains (see Fig. 4.21). Regarding to STAT2, 55 residues have a high frequency where 19 of them are involved with any type of interaction. The residues are Arg-796, Arg-88, Asp-77, Asp-794, Asp-850, Asp-93, Glu-40, Glu-715, Glu-722, Glu-79, Glu-801, Glu-804, Glu-814, His-85, Leu-684, Leu-691, Leu-727, Leu-81, and Lys-89 that have been located in the N-terminal domain (see Fig. 4.21). Studies of hot-spot in the binding sites of protein-protein interface have located a particular enrichment of Trp, Try and Arg, as well as a high presence of polar residues [99]. In contrast, hydrophobic residues as Val and Leu are associated to interfaces largely hydrophobic and nonpolar surface areas [99]. In our case, Arg and Glu are hydrophilic residues with largest presence in the interface of NS5 and STAT2, respectively, besides hydrophilic residues as Lys, Pro, and Asp are also displayed in the interface of both proteins. Other studies have concluded that residues as Trp, Met, Try, Phe, Cys, and Ile are frequently in the binding interfaces. Besides, residues as Tyr, Trp, His, and Cys have been traced in high-affinity interfaces in comparison with low-affinity interfaces [100]. His and Trp are present in the interface of NS5-STAT2 complex, hence the interface shows a high-affinity. However, residue as Lys are located in low-affinity interface. In our outcomes, this latter one have been found in the interface of NS5-STAT2 complex which will counteract a possible high-affinity in the NS5-STAT2 complex [100]. As was mentioned before, NS5-STAT2 is a heterocomplex which has differences properties associated to amino acid composition, contact sites, and interface area [101]. Hence, in studies of homo- and hetero- complexes has revealed that amino acid interface composition is different between them. To heterocomplex residues that interact as Leu, Val, Ile, Arg, Tyr, Trp, Met, and Phe have a great prevalence and propensity to be in the interface. These types of residues have also found in our outcomes excepting Ile, Met and Phe [101]. Predominant residues in our results have also been associated with interaction between chains in the complex, hence Asp and Glu are implicated with interaction through their main chain. However, Lys involves the contact with the backbone and Pro through the side chain [101].

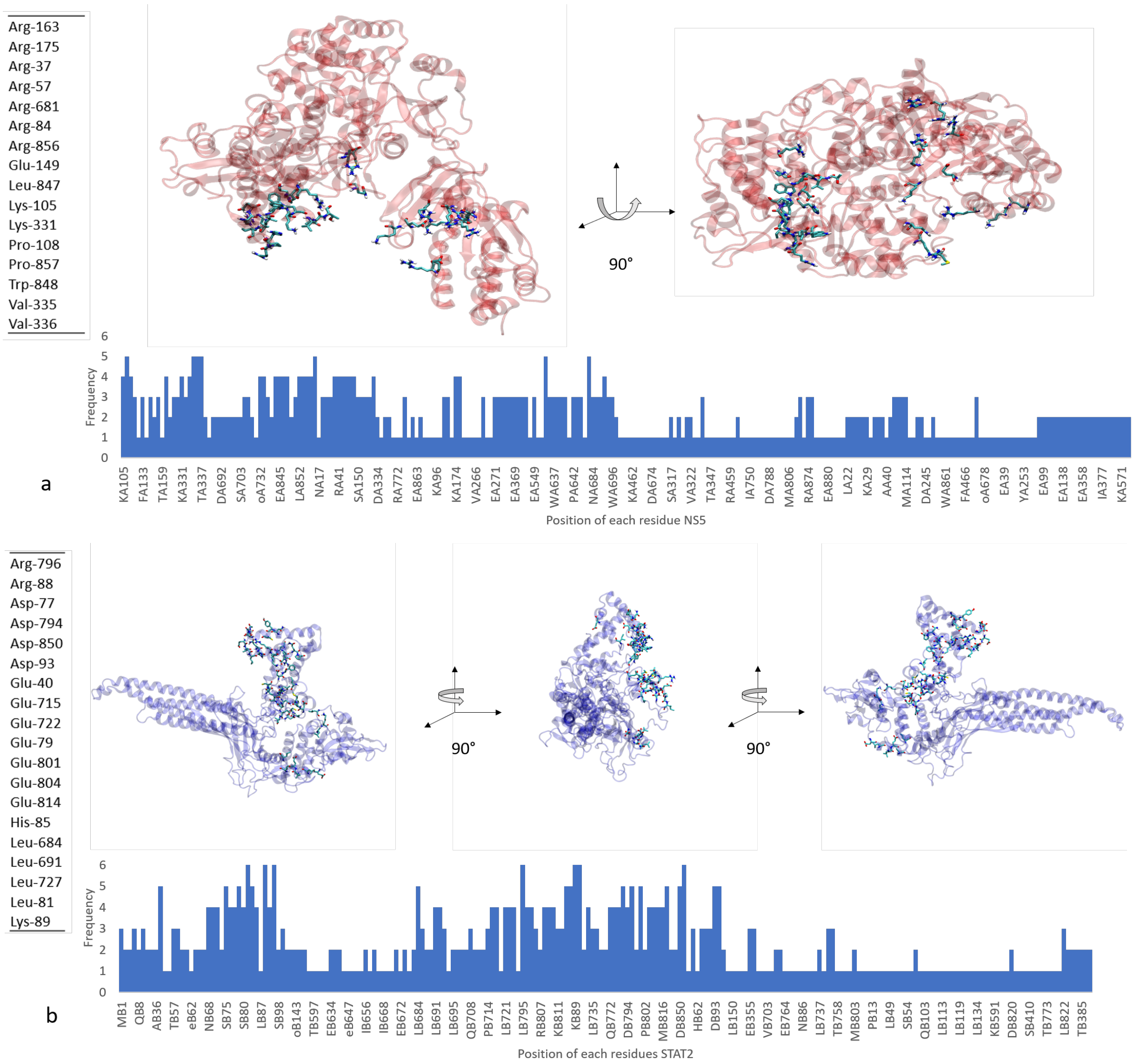


Figure 4.21. Frequency of residues in the docked complex that have interactions in the interface. High frequency of residues in (a) NS5 and (b) STAT2 that showed any type of interaction such as electrostatic,hydrophobic, salt bridges or ionic interaction.

Chapter 5

Conclusions

The computational approach has permitted to analyze the different structural and dynamics features of NS5, STAT2 and the interaction of NS5-STAT2 complex. Our study has been undertaken in two directions; the first through analysis of amino acid sequence of NS5 and STAT2, and the second corresponding to the atomic coordinates of NS5, STAT2 and NS5-STAT2 complex. The prediction of protein disorder based on sequences was performed. STAT2 has shown regions with greater disorder than NS5, and are located in the coiled-coil, DNA-binding and Transactivation domains. The three-dimensional model of STAT2 generated by I-TASSER was validated through several computational tools. This model displays a high correlation with experimental fragments of STAT2. The MD simulations were individually performed to STAT and NS5 showing a high stability during the period of simulation. The atomic fluctuation and solvent access of both proteins displayed correspondence with the disorder prediction and shape. The interaction between NS5 and STAT2 was reached through docking approach. Several models were analyzed determining that three docked complexes provided by ClusPro have interaction among the domains of N-terminal domain from STAT2 and Mtase-Thumb domains from NS5. The interaction between these domains has been confirmed through contact mapping and electrostatic analysis. MD simulations for the three docked complexes have suggested that their behavior is affected for each other in interactions that are favorable to binding in the interface since it has shown a $\Delta G_{Bind}<0$. The best docked complex have a ΔG_{Bind} of -4.30

kcal·mol^{-1}. Futhermore, the NS5-STAT2 docked complex has revealed the key interacting residues are stabilized by electrostatic interaction, hydrophobic interaction, salt bridges and ionic interaction. Therefore, it suggests that the interaction between these proteins is focused on the three domains (N-terminal/Mtase and N-terminal/Thumb) which are ordered regions enriched with polar residues. In addition, this study sheds light in the interaction of NS5-STAT2 as support of the experimental studies and in the development of drugs against ZIKV NS5.

Bibliography

[1] L. Jian, G. Hansen, C. Nitsche, C. D. Klein, L. Zhang, and R. Hilgenfeld, “Crystal structure of zika virus ns2b-ns3 protease in complex with a boronate inhibitor,” *Science*, vol. 353, no. 6298, pp. 503–505, 2016.

[2] C. F. Baez, V. A. Barel, A. M. De Souza, C. R. Rodrigues, R. B. Varella, and N. Cirauqui, “Analysis of worldwide sequence mutations in Zika virus proteins E, NS1, NS3 and NS5 from a structural point of view,” *Molecular BioSystems*, vol. 13, no. 1, pp. 122–131, 2017.

[3] B. D. Cox, R. A. Stanton, and R. F. Schinazi, “Predicting Zika virus structural biology: Challenges and opportunities for intervention,” *Antiviral Chemistry and Chemotherapy*, vol. 24, no. 3-4, pp. 118–126, 2015.

[4] C. H. Chuang, S. J. Chiou, T. L. Cheng, and Y. T. Wang, “A molecular dynamics simulation study decodes the Zika virus NS5 methyltransferase bound to SAH and RNA analogue,” *Scientific Reports*, vol. 8, no. 1, pp. 1–9, 2018.

[5] B. Wang, X.-F. Tan, S. Thurmond, Z.-M. Zhang, A. Lin, R. Hai, and J. Song, “The structure of Zika virus NS5 reveals a conserved domain conformation,” *Nature Communications*, vol. 8, p. 14763, 2017.

[6] N. J. da Fonseca, M. Q. Lima Afonso, N. G. Pedersolli, L. C. de Oliveira, D. S. Andrade, and L. Bleicher, “Sequence, structure and function relationships in flaviviruses as assessed by evolutive aspects of its conserved non-structural protein domains,” *Biochemical and Biophysical Research Communications*, pp. 1–7, 2016.

[7] B. Zhao, G. Yi, F. Du, Y.-c. Chuang, R. C. Vaughan, B. Sankaran, C. C. Kao, and P. Li, “Structure and function of the Zika virus full-length NS5 protein,” *Nature Communications*, vol. 8, pp. 1–9, 2017.

[8] A. Grant, S. S. Ponia, S. Tripathi, V. Balasubramaniam, L. Miorin, M. Sourisseau, M. C. Schwarz, M. P. Sánchez-Seco, M. J. Evans, S. M. Best, and A. García-Sastre, “Zika Virus Targets Human STAT2 to Inhibit Type i Interferon Signaling,” *Cell Host and Microbe*, vol. 19, no. 6, pp. 882–890, 2016.

[9] K.-i. Arimoto, S. Löchte, S. A. Stoner, C. Burkart, Y. Zhang, S. Miyauchi, S. Wilmes, J.-B. Fan, J. J. Heinisch, Z. Li, M. Yan, S. Pellegrini, F. Colland, J. Piehler, and D.-E. Zhang, “STAT2 is an essential adaptor in USP18-mediated suppression of type I interferon signaling,” *Nature Structural & Molecular Biology*, vol. 24, no. 3, pp. 279–289, 2017.

[10] J. R. Bowen, K. M. Quicke, M. S. Maddur, J. T. O’Neal, C. E. McDonald, N. B. Fedorova, V. Puri, R. S. Shabman, B. Pulendran, and M. S. Suthar, “Zika Virus Antagonizes Type I Interferon Responses during Infection of Human Dendritic Cells,” *PLOS Pathogens*, vol. 13, no. 2, p. e1006164, 2017.

[11] A. Kumar, S. Hou, A. M. Airo, D. Limonta, V. Mancinelli, W. Branton, C. Power, and T. C. Hobman, “Zika virus inhibits type-I interferon production and downstream signaling.,” *EMBO reports*, vol. 17, no. 12, pp. 487–524, 2016.

[12] A. S. Godoy, G. M. A. Lima, K. I. Z. Oliveira, N. U. Torres, F. V. Maluf, R. V. C. Guido, and G. Oliva, “Crystal structure of Zika virus NS5 RNA-dependent RNA polymerase,” *Nature Communications*, vol. 8, p. 14764, 2017.

[13] C. P. Lim and X. Cao, “Structure, function, and regulation of STAT proteins,” *Molecular BioSystems*, vol. 2, no. 11, pp. 536–550, 2006.

[14] J. Ho, C. Pelzel, A. Begitt, M. Mee, H. M. Elsheikha, D. J. Scott, and U. Vinkemeier, “STAT2 Is a Pervasive Cytokine Regulator due to Its Inhibition of STAT1 in Multiple Signaling Pathways,” *PLoS Biology*, vol. 14, no. 10, pp. 1–27, 2016.

[15] N. Raftery and N. J. Stevenson, “Advances in anti-viral immune defence: revealing the importance of the IFN JAK/STAT pathway,” *Cellular and Molecular Life Sciences*, pp. 1–11, 2017.

[16] K. Blaszczyk, H. Nowicka, K. Kostyrko, A. Antonczyk, J. Wesoly, and H. A. Bluyssen, “The unique role of STAT2 in constitutive and IFN-induced transcription and antiviral responses,” *Cytokine and Growth Factor Reviews*, vol. 29, pp. 71–81, 2016.

[17] K. E. Reed, A. E. Gorbalenya, and C. M. Rice, “The NS5A/NS5 Proteins of Viruses from Three Genera of the Family Flaviviridae Are Phosphorylated by Associated Serine/Threonine Kinases,” *Journal of Virology*, vol. 72, no. 7, pp. 6199–6206, 1998.

[18] F. Plattner and J. A. Bibb, *Serine and Threonine Phosphorylation.* Elsevier Inc., eighth edition ed., 2012.

[19] M. Johansson, A. J. Brooks, D. A. Jans, and S. G. Vasudevan, “A small region of the dengue virus-encoded RNA-dependent RNA polymerase, NS5, confers interaction with both the nuclear transport receptor importin-β and the viral helicase, NS3,” *Journal of General Virology*, vol. 82, no. 4, pp. 735–745, 2001.

[20] R. Kapoor, Mini and Zhang, Luwen and Ramachandra, Muralidhara and Kusukawa, Jingo and Ebner, Kurt E and Padmanabhan, “Association between NS3 and NS5 Protein of Dengue Virus Type in the Putative RNA Replicase Is Linked to Differential Phosphorylation of NS5,” *Journal of Biological Chemistry*, vol. 32, pp. 19100–19106, 1995.

[21] R. Yan, S. Qureshi, Z. Zhong, Z. Wen, and J. E. Darnell, "The genomic structure of the STAT genes: Multiple exons in coincident sites in stat1 and stat2," *Nucleic Acids Research*, vol. 23, no. 3, pp. 459–463, 1995.

[22] S. Rengachari, S. Groiss, E. Caron, and N. Grandvaux, "Structural basis of STAT2 recognition by IRF9 reveals molecular insights into ISGF3 function," 2017.

[23] C. M. Horvath, Z. Wen, and J. E. D. Jr, "recognition suggests a novel DNA-binding domain . A STAT protein domain that determines DNA sequence recognition suggests a novel DNA-binding domain," pp. 984–994, 1995.

[24] S. H. Mahboobi, A. A. Javanpour, and M. R. K. Mofrad, "The interaction of RNA helicase DDX3 with HIV-1 Rev-CRM1-RanGTP complex during the HIV replication cycle," *PLoS ONE*, vol. 10, no. 2, 2015.

[25] I. Bhutani, S. Loharch, P. Gupta, R. Madathil, and R. Parkesh, "Structure, dynamics, and interaction of Mycobacterium tuberculosis (Mtb) DprE1 and DprE2 examined by molecular modeling, simulation, and electrostatic studies," *PLoS ONE*, vol. 10, no. 3, pp. 1–31, 2015.

[26] A. Fiser, "Template-Based Protein Structure Modeling," vol. 673, pp. 1–20, 2010.

[27] T. Schwede, A. Sali, and N. Eswar, "Protein Structure Modeling," in *Computational Structural Biology: methods and applications* (T. Schwede and M. C. Peitsch, eds.), ch. 1, pp. 1–33, Danvers: World Scientific Publishing Co. Pte. Ltd, 2008.

[28] A. Hospital, J. R. Goñi, M. Orozco, and J. Gelpi, "Molecular dynamics simulations: Advances and applications," *Advances and Applications in Bioinformatics and Chemistry*, vol. 8, pp. 37–47, 2015.

[29] J. L. Klepeis, K. Lindorff-Larsen, R. O. Dror, and D. E. Shaw, "Long-timescale molecular dynamics simulations of protein structure and function," *Current Opinion in Structural Biology*, vol. 19, no. 2, pp. 120–127, 2009.

[30] D. Vlachakis, E. Bencurova, N. Papangelopoulos, and S. Kossida, *Current state-of-the-art molecular dynamics methods and applications*, vol. 94. Elsevier Inc., 1 ed., 2014.

[31] M. Karplus and J. A. Mccammon, “Corrigenda: Molecular dynamics simulations of biomolecules,” *Nature Structural Biology*, vol. 9, no. 10, pp. 788–788, 2002.

[32] C. Sagui and T. A. Darden, “MOLECULAR DYNAMICS SIMULATIONS OF BIOMOLECULES: Long-Range Electrostatic Effects,” *Annual Review of Biophysics and Biomolecular Structure*, vol. 28, no. 1, pp. 155–179, 1999.

[33] M. Karplus and G. a. Petsko, “Molecular dynamics simulations in biology.,” *Nature*, vol. 347, no. 6294, pp. 631–639, 1990.

[34] T. and others Duan, Yong and Wu, Chun and Chowdhury, Shibasish and Lee, Mathew C and Xiong, Guoming and Zhang, Wei and Yang, Rong and Cieplak, Piotr and Luo, Ray and Lee, Taisung and othersDuan, Yong and Wu, Chun and Chowdhury, Shibasish and Lee, Mathew C and Xiong, Gu, “A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations,” *Journal of computational chemistry*, vol. 24, no. 16, pp. 1999–2012, 2003.

[35] W. Frenkel, Daan and Smit, Berend and Tobochnik, Jan and McKay, Susan R and Christian, “Understanding Molecular Simulation,” *Computers in Physics*, vol. 11, no. 4, pp. 351–354, 1997.

[36] J. P. Mithen, “Molecular dynamics simulations of the equilibrium dynamics of non-ideal plasmas,” *Computing*, 2012.

[37] M. Abraham, B. Hess, D. van der Spoel, and E. Lindahl, “GROMACS User Manual version 5.0.7,” *Www.Gromacs.Org*, 2015.

[38] P. H. Hünenberger, “Thermostat algorithms for molecular dynamics simulations,” *Advances in Polymer Science*, vol. 173, pp. 105–147, 2005.

[39] D. J. Evans and B. L. Holian, “The Nose-Hoover thermostat,” *The Journal of Chemical Physics*, vol. 83, no. 8, pp. 4069–4074, 1985.

[40] H. J. Berendsen, J. P. Postma, W. F. Van Gunsteren, A. Dinola, and J. R. Haak, “Molecular dynamics with coupling to an external bath,” *The Journal of Chemical Physics*, vol. 81, no. 8, pp. 3684–3690, 1984.

[41] I. Halperin, B. Ma, H. Wolfson, and R. Nussinov, “Principles of docking: An overview of search algorithms and a guide to scoring functions,” *Proteins: Structure, Function and Genetics*, vol. 47, no. 4, pp. 409–443, 2002.

[42] I. D. Brooijmans, Natasja and Kuntz, “Molecular recognition and docking algorithms,” *Annual review of biophysics and biomolecular structure*, vol. 32, no. 1, pp. 335–373, 2003.

[43] G. R. Smith and M. J. Sternberg, “Prediction of protein-protein interactions by docking methods,” *Current Opinion in Structural Biology*, vol. 12, no. 1, pp. 28–35, 2002.

[44] F. M. Ytreberg, R. H. Swendsen, and D. M. Zuckerman, “Comparison of free energy methods for molecular systems,” *Journal of Chemical Physics*, vol. 125, no. 18, pp. 1–11, 2006.

[45] K. Tamura, G. Stecher, D. Peterson, A. Filipski, and S. Kumar, “MEGA6: Molecular evolutionary genetics analysis version 6.0,” *Molecular Biology and Evolution*, vol. 30, no. 12, pp. 2725–2729, 2013.

[46] D. T. Jones and D. Cozzetto, “DISOPRED3: Precise disordered region predictions with annotated protein-binding activity,” *Bioinformatics*, vol. 31, no. 6, pp. 857–863, 2015.

[47] M. Källberg, H. Wang, S. Wang, J. Peng, Z. Wang, H. Lu, and J. Xu, “Template-based protein structure modeling using the RaptorX web server,” *Nature Protocols*, vol. 7, no. 8, pp. 1511–1522, 2012.

[48] J. Ma, S. Wang, Z. Wang, and J. Xu, “Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning,” *Bioinformatics*, vol. 31, no. 21, pp. 3506–3513, 2014.

[49] S. Wang, W. Li, S. Liu, and J. Xu, “RaptorX-Property: a web server for protein structure property prediction,” *Nucleic acids research*, vol. 44, no. W1, pp. W430–W435, 2016.

[50] J.-F. J.-f. Gibrat, J. Garnier, and B. Robson, “[32] GOR method for predicting protein secondary structure from amino acid sequence,” *Computer Methods for Macromolecular Sequence Analysis*, vol. Volume 266, no. 1995, pp. 540–553, 1996.

[51] A. Roy, A. Kucukural, and Y. Zhang, “I-TASSER: a unified platform for automated protein structure and function prediction,” *Nature Protocols*, vol. 5, no. 4, pp. 725–738, 2010.

[52] K. L.A., M. S., Y. C., W. M., and S. M, “The Phyre2 web portal for protein modelling, prediction, and analysis,” *Nature Protocols*, vol. 10, no. 6, pp. 845–858, 2015.

[53] D. W. A. Buchan, F. Minneci, T. C. O. Nugent, K. Bryson, and D. T. Jones, “Scalable web services for the PSIPRED Protein Analysis Workbench.,” *Nucleic acids research*, vol. 41, no. Web Server issue, pp. 349–357, 2013.

[54] M. Remmert, A. Biegert, A. Hauser, and J. Söding, “HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment,” *Nature Methods*, vol. 9, no. 2, pp. 173–175, 2012.

[55] R. Lüthy, J. U. Bowie, and D. Eisenberg, “Assessment of protein models with three-dimensional profiles,” *Nature*, vol. 356, no. 6364, pp. 83–85, 1992.

[56] C. Colovos and T. O. Yeates, “Verification of protein structures: Patterns of non-bonded atomic interactions,” *Protein Science*, vol. 2, no. 9, pp. 1511–1519, 1993.

[57] L. Willard, A. Ranjan, H. Zhang, H. Monzavi, R. F. Boyko, B. D. Sykes, and D. S. Wishart, “VADAR: A web server for quantitative evaluation of protein structure quality,” *Nucleic Acids Research*, vol. 31, no. 13, pp. 3316–3319, 2003.

[58] Y. Zhang and J. Skolnick, “TM-align: A protein structure alignment algorithm based on the TM-score,” *Nucleic Acids Research*, vol. 33, no. 7, pp. 2302–2309, 2005.

[59] M. L. Jorgensen, William L and Chandrasekhar, Jayaraman and Madura, Jeffry D and Impey, Roger W and Klein, “Comparison of simple potential functions for simulating liquid water,” *The Journal of Chemical Physics*, vol. 79, no. 2, pp. 926–935, 1983.

[60] S. Nosé, “A unified formulation of the constant temperature molecular dynamics methods,” *The Journal of Chemical Physics*, vol. 81, no. 1, pp. 511–519, 1984.

[61] M. Parrinello and A. Rahman, “Polymorphic transitions in single crystals: A new molecular dynamics method,” *Journal of Applied Physics*, vol. 52, no. 12, pp. 7182–7190, 1981.

[62] B. Hess, H. Bekker, H. J. Berendsen, and J. G. Fraaije, “LINCS: A Linear Constraint Solver for molecular simulations,” *Journal of Computational Chemistry*, vol. 18, no. 12, pp. 1463–1472, 1997.

[63] X. Daura, K. Gademann, B. Jaun, D. Seebach, W. F. van Gunsteren, and A. E. Mark, “Peptide Folding: When Simulation Meets Experiment,” *Angewandte Chemie International Edition*, vol. 38, no. 1-2, pp. 236–240, 1999.

[64] D. Kozakov, D. R. Hall, B. Xia, K. A. Porter, D. Padhorny, C. Yueh, D. Beglov, and S. Vajda, “The ClusPro web server for proteinprotein docking,” *Nature Protocols*, vol. 12, no. 2, pp. 255–278, 2017.

[65] B. Jiménez-García, C. Pons, and J. Fernández-Recio, "pyDockWEB: A web server for rigid-body protein-protein docking using electrostatics and desolvation scoring," *Bioinformatics*, vol. 29, no. 13, pp. 1698–1699, 2013.

[66] A. Sukhwal and R. Sowdhamini, "Oligomerisation status and evolutionary conservation of interfaces of protein structural domain superfamilies," *Molecular BioSystems*, vol. 9, no. 7, pp. 1652–1661, 2013.

[67] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano, "The FoldX web server: An online force field," *Nucleic Acids Research*, vol. 33, no. SUPPL. 2, pp. 382–388, 2005.

[68] E. Krissinel and K. Henrick, "Inference of Macromolecular Assemblies from Crystalline State," *Journal of Molecular Biology*, vol. 372, no. 3, pp. 774–797, 2007.

[69] K. G. Tina, R. Bhadra, and N. Srinivasan, "PIC: Protein Interactions Calculator," *Nucleic Acids Research*, vol. 35, no. SUPPL.2, pp. 473–476, 2007.

[70] A. Vangone, R. Spinelli, V. Scarano, L. Cavallo, and R. Oliva, "COCOMAPS: A web application to analyze and visualize contacts at the interface of biomolecular complexes," *Bioinformatics*, vol. 27, no. 20, pp. 2915–2916, 2011.

[71] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon, "Electrostatics of nanosystems: Application to microtubules and the ribosome," *Proceedings of the National Academy of Sciences*, vol. 98, no. 18, pp. 10037–10041, 2001.

[72] T. J. Dolinsky, P. Czodrowski, H. Li, J. E. Nielsen, J. H. Jensen, G. Klebe, and N. A. Baker, "PDB2PQR: Expanding and upgrading automated preparation of biomolecular structures for molecular simulations," *Nucleic Acids Research*, vol. 35, no. SUPPL.2, pp. 522–525, 2007.

[73] E. Jurrus, D. Engel, K. Star, K. Monson, J. Brandi, L. E. Felberg, D. H. Brookes, L. Wilson, J. Chen, K. Liles, M. Chun, P. Li, D. W. Gohara, T. Dolinsky, R. Konecny, D. R. Koes, J. E. Nielsen, T. Head-Gordon, W. Geng, R. Krasny,

G. W. Wei, M. J. Holst, J. A. McCammon, and N. A. Baker, "Improvements to the APBS biomolecular solvation software suite," *Protein Science*, vol. 27, no. 1, pp. 112–128, 2018.

[74] J. D. Atkins, S. Y. Boateng, T. Sorensen, and L. J. McGuffin, "Disorder prediction methods, their applicability to different protein targets and their usefulness for guiding experimental studies," *International Journal of Molecular Sciences*, vol. 16, no. 8, pp. 19040–19054, 2015.

[75] P. Lieutaud, F. Ferron, A. V. Uversky, L. Kurgan, V. N. Uversky, and S. Longhi, "How disordered is my protein and what is its disorder for? A guide through the dark side of the protein universe," *Intrinsically Disordered Proteins*, vol. 4, no. 1, p. e1259708, 2016.

[76] B. Mészáros, P. Tompa, I. Simon, and Z. Dosztányi, "Molecular Principles of the Interactions of Disordered Proteins," *Journal of Molecular Biology*, vol. 372, no. 2, pp. 549–561, 2007.

[77] Y. Y. Ji and Y. Q. Li, "The role of secondary structure in protein structure selection," *European Physical Journal E*, vol. 32, no. 1, pp. 103–107, 2010.

[78] S. C. Kwok, C. T. Mant, and R. S. Hodges, "Importance of secondary structural specificity determinants in protein folding: insertion of a native β-sheet sequence into an α-helical coiled-coil.," *Protein science*, vol. 11, no. 6, pp. 1519–31, 2002.

[79] J. Ma and S. Wang, "AcconPred: Predicting solvent accessibility and contact number simultaneously by a multitask learning framework under the conditional neural fields model," *BioMed Research International*, vol. 2015, 2015.

[80] L. Palmieri, M. Federico, M. Leoncini, and M. Montangero, "A high performing tool for residue solvent accessibility prediction," in *International Conference on Information Technology in Bio-and Medical Informatics*, pp. 138–152, Springer, 2011.

[81] H. Chen and H. X. Zhou, “Prediction of solvent accessibility and sites of deleterious mutations from protein sequence,” *Nucleic Acids Research*, vol. 33, no. 10, pp. 3193–3199, 2005.

[82] C. Chothia, W. Ramsay, R. Foster, and C. Ingold, “Solvent accessibility, protein surfaces, and protein folding,” *Biophysical Journal*, vol. 32, no. 1, pp. 35–44, 1980.

[83] T. Sikosek and H. S. Chan, “Biophysics of protein evolution and evolutionary protein biophysics,” *Journal of The Royal Society Interface*, vol. 11, no. 100, pp. 20140419–20140419, 2014.

[84] M. C. Deller, L. Kong, and B. Rupp, “Protein stability: A crystallographer’s perspective,” *Acta Crystallographica Section:F Structural Biology Communications*, vol. 72, pp. 72–95, 2016.

[85] I. S. Moreira, J. M. Martins, J. T. Coimbra, M. J. Ramos, and P. A. Fernandes, “A new scoring function for protein-protein docking that identifies native structures with unprecedented accuracy,” *Physical Chemistry Chemical Physics*, vol. 17, no. 4, pp. 2378–2387, 2015.

[86] O. Mathew and R. Sowdhamini, “PIMA: Protein-Protein Interactions in Macromolecular Assembly - a web server for its Analysis and Visualization,” *Bioinformation*, vol. 12, no. 1, pp. 9–11, 2016.

[87] O. K. Mathew and R. Sowdhamini, “PIMADb: A database of protein-protein interactions in huge macromolecular assemblies,” *Bioinformatics and Biology Insights*, vol. 10, pp. 105–109, 2016.

[88] A. H. Elcock, D. Sept, and J. A. Mccammon, “Computer Simulation of Protein Protein Interactions,” *J. Phys. Chem. B*, pp. 1504–1518, 2001.

[89] K. O., T. N., and G. A., “Characterization and prediction of protein interfaces to infer protein-protein interaction networks,” *Current Pharmaceutical Biotechnology*, vol. 9, no. 2, pp. 67–76, 2008.

[90] M. C. Smith and J. E. Gestwicki, “Features of protein–protein interactions that translate into potent inhibitors: topology, surface area and affinity,” *Expert reviews in molecular medicine*, vol. 14, 2012.

[91] D. H. Williams, E. Stephens, D. P. O’Brien, and M. Zhou, “Understanding noncovalent interactions: Ligand binding energy and catalytic efficiency from ligand-induced reductions in motion within receptors and enzymes,” *Angewandte Chemie - International Edition*, vol. 43, no. 48, pp. 6596–6616, 2004.

[92] E. Zhang, Zhe and Witham, Shawn and Alexov, “On the role of electrostatics on protein-protein interactions,” *Physical biology*, vol. 8, no. 3, p. 035001, 2011.

[93] F. B. Sheinerman, R. Norel, and B. Honig, “Electrostatic aspects of protein-protein interactions,” *Current Opinion in Structural Biology*, vol. 10, no. 2, pp. 153–159, 2000.

[94] P. J. Kundrotas and E. Alexov, “Electrostatic properties of protein-protein complexes,” *Biophysical Journal*, vol. 91, no. 5, pp. 1724–1736, 2006.

[95] P. W. Snyder, M. R. Lockett, D. T. Moustakas, and G. M. Whitesides, “Is it the shape of the cavity, or the shape of the water in the cavity?,” *European Physical Journal: Special Topics*, vol. 223, no. 5, pp. 853–891, 2014.

[96] T. Vreven, H. Hwang, B. G. Pierce, and Z. Weng, “Prediction of protein-protein binding free energies,” *Protein Science*, vol. 21, no. 3, pp. 396–404, 2012.

[97] X. H. Ma, C. X. Wang, C. H. Li, and W. Z. Chen, “A fast empirical approach to binding free energy calculations based on protein interface information.,” *Protein engineering*, vol. 15, no. 8, pp. 677–681, 2002.

[98] A. F. Brito and J. W. Pinney, “Protein-protein interactions in virus-host systems,” *Frontiers in Microbiology*, vol. 8, no. AUG, pp. 1–11, 2017.

[99] B. Ma, T. Elkayam, H. Wolfson, and R. Nussinov, "Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces," *Proceedings of the National Academy of Sciences*, vol. 100, no. 10, pp. 5772–5777, 2003.

[100] A. Erijman, E. Rosenthal, and J. M. Shifman, "How structure defines affinity in protein-protein interactions," *PLoS ONE*, vol. 9, no. 10, 2014.

[101] D. Talavera, D. L. Robertson, and S. C. Lovell, "Characterization of protein-protein interaction interfaces from a single species," *PLoS ONE*, vol. 6, no. 6, 2011.