École Normale Supérieure de Cachan
École Doctorale de Sciences Pratiques

Universidade Federal do Rio de Janeiro
Instituto de Biofísica Carlos Chagas Filho

Joint-PhD thesis:

Oncogenic Mechanisms of Activation and Resistance of the type III Receptor Tyrosine Kinase Family

PhD student: Priscila da Silva Figueiredo Celestino Gomes

Advisors: Prof. Pedro Geraldo Pascutti

Dr. Luba Tchertanov

June 2015

Acknowledgments

First of all, I would like to thank my family, specially my beloved parents, for all the financial and emotional support conceded, not only during the past four years, but all the path that I had to traverse to get to this point of my academic life. I will never be able to repay you for all the sacrifices done and for your dedication.

I would like to thank my loved husband, Diego, for being so patient and kind during all the difficult times and for supporting and loving me unconditionally. For putting up with my mood and difficult temper (sometimes?) during the thesis writing and for preparing me delicious dinners almost every day! For the midnight discussions, sleepless nights. For believing in me even when I couldn't do it myself.

I find it difficult to express in words all the things I would like to say to thank my supervisors. Pedro, thank you for welcoming me in your lab and supporting my work since I was an undergrad student. Thank you for supporting and stimulating my scientific exchange experience in France. Without your support, I could never have done that in the first place. Thank you for being so kind and for giving us freedom in the scientific projects, respecting everyone`s ideas and thoughts. Finally, I would like to thank you for the fruitful discussions and life-saving brainstorms. Luba, thank you for welcoming me (three) times in your group in France and giving me the opportunity to leave my comfort zone to throw myself into new challenges. My experience in France has changed my perspective in many ways, personal included. Thank you for the long discussions in your office and to pushing me into a better version of myself academically. Thank you for your patience and way of treating things so gracefully.

Last but not least, I would like to thank all the friends that accompanied me in this (long) journey. Beginning with the Brazilian side, I would like to thank all the lab co-workers that contributed somehow to enrich my work at the discussions in the lab: Pedro Torres, Marcelo, Tácio, Reinaldo, Moema, Rosemberg, among others. A special thanks to Pedro Torres, for being the very definition of "comrade" and listening all of my misery and misfortune events in a daily basis. A special thanks to my girlfriends Raysa, Bia and Manu for their sincere friendship, being always at my side, always so comprehensive and thoughtful. For the French side, I also would like to thank all my lab co-workers during the (almost) three years spent at Cachan: Elodie, Isaure, Rohit, Florent, Nolan, Nicolas and Yann, among others. Elodie, thank you for the enriching discussions and for accompanying me at the first and critical year of my PhD; thank you for reminding me that we can find always a bright side in every result and thank you for being a friend when I needed. I would like also to leave a special thanks to the boys, Nicolas, Nolan, Florent and Yann, for welcoming me so openly into your convivial meetings. I know it is hard to achieve that being a foreigner! Thank you Nolan for cheering me up all the time with your sunny personality; thank you Nico for being the best *stagiaire* ever and participating lively in my work in the last six months of my stay. *Enfin*, thank you all for the great times we had together, it meant a lot to me. I could not forget to mention the Brazilians I have the chance to meet during my stay in France, thank you all for the company, and shared weekends together: Alexandre, Livia, Leticia, Larissa and Isabella. Special thanks to Isabella for having her house always with open doors (especially kitchen!) and for her honest friendship.

"To raise new questions, new possibilities, to regard old problems from a new angle, requires creative imagination and marks real advance in science"

-Albert Einstein

"There is a theory which states that if ever anyone discovers exactly what the Universe is for and why it is here, it will instantly disappear and be replaced by something even more bizarre and inexplicable.
There is another theory which states that this has already happened."
-Douglas Adams, *The Restaurant at the End of the Universe*.

Summary

Acknowledgments II
Summary IV
List of tables VII
List of figures IX
Abbreviations and acronyms XVII
Resumo XX
Abstract XXI
Résumé XXII
Chapter 1: Introduction 1
1. Tyrosine kinase receptors 1
2. Type III RTK subfamily 2
2.1. CSF-1 and its receptor CSF-1R 5
2.1.1. Cell signaling pathways activated by CSF-1R 6
2.1.2. The CSF-1 role in cancer 8
2.2. KIT receptor and hotspot mutations 9
3. Allosteric regulation of RTKs 20
4. Molecular modeling of bio-macromolecules 23
4.1. Protein structure and interactions 24
4.2. Molecular mechanics, Quantum mechanics and Force fields 25
4.3. Minimization methods for geometry optimization 28
5. Protein structure prediction 30
5.1. Secondary structure prediction methods 30
5.2. Tridimensional protein structure prediction 33
5.2.1. Comparative modeling 34
5.2.2. Template-free protein modeling 38
6. Molecular dynamics simulations 39
6.1. General MD protocol 42
6.2. Analysis of MD trajectories 48
6.2.1. Root Mean Square Deviation (RMSD) and Root Mean Square Fluctuation (RMSF) 48
6.2.2. Convergence analysis 48
6.2.3. Secondary structure 49

6.2.4. Principal Components Analysis 49
6.2.5. Free energy of binding by the Molecular Mechanics (Poisson-Boltzmann /Generalized Born) Surface Area approach 50
7. Normal modes analysis 52
8. Modular network analysis with MONETA 53
9. Molecular docking 57
9.1. Receptor flexibility 58
9.2. Ligand sampling 59
9.3. Scoring functions 60
Goals 62
Chapter 2: Methodology 64
1. Structural and dynamical study of wild-type and mutant forms of CSF-1R 64
1.1. Secondary structure prediction of the JMR 64
1.2. Electrostatic potential surface 64
1.3. Preparation of initial coordinates 64
1.4. Molecular dynamics simulations 65
1.5. Energy analysis 67
1.6. Normal modes analysis (NMA) 67
1.7. Principal Component Analysis (PCA) 69
1.8. Analysis of intramolecular communication 69
2. CSF-1R and KIT receptors complexed with imatinib 70
2.1. Preparation of initial coordinates 70
2.2. Molecular docking 72
2.3. Molecular dynamics simulations of the imatinib-target complexes 73
2.4. Energy analysis 75
2.5. MD simulations for KIT forms containing the truncated JMR 76
Chapter 3: Results 77
1. Structural and dynamic study of the wild-type CSF-1R and the D802V mutant: comparison with the effects observed in KIT^{D816V} 77
2. Imatinib binding mode to the CSF-1R and KIT receptors in their native and mutated forms 101
Discussion 132
Conclusions 137
Appendix 138
1. Modeling of the full length structure for the CSF-1R`s cytoplasmic domain by prediction of the KID structure 138
1.1. Methodology 138
1.1.1. Bioinformatics analysis 138

1.1.2. Modeling protocols 139
1.2. Preliminary results 144
2. Simulations parameters file 152
2.1. Docking 152
2.2. MD simulations 153
2.3. Energy analysis 184
Bibliography 186

List of tables

Table 1: Commonly Used Tools and Services for Protein Structure Modeling and Prediction. Adapted from: (SCHWEDE, 2013) 34

Table 2: Parameters used in the convergence analysis of the native CSF-1R (WT) and its mutant (D802V) MD trajectories. 88

Table 3: H-bonds stabilized the inactive conformation in CSF-1R^{WT} and CSF1R^{MU}. Residues involved in H-bonding and the H-bond occurrences (in %) are computed and averaged over MD simulations. Occurrences showed a major difference are denoted by an asterisk. 91

Table 4: Quantitative analysis of the communication network pattern among the different MD replicas. MD1, MD2 and MD12 are the two separate and merged trajectories respectively. 96

*Table 5: Score and RMSD of the best poses generated by the Induced Fit protocol. Maestro outputs the energy values in two scores: GlideScore and Induced Fit score (IFD). RMSD values were calculated using the crystal structure of inactive KIT complexed with imatinib (PDB ID: 1T46).*IFD score of KIT^{V560G} is extremely low due to the constriction used in the docking simulations (see Methods section).* 106

Table 6: H-bond occurrences related to the P-loop residue D778 from CSF-1R. The values correspond to the average from both MD simulation replicas. 110

Table 7: Hydrogen bonds (H-bonds) occurrences between the targets and imatinib, averaged over the two MD replicas. The atom pairs for donor and acceptor interactions are depicted in the table. Imatinib atoms participating in the interaction are represented in the 2D representation of the inhibitor's structure. 112

Table 8: CSF-1R residues contribution to the final binding energy of imatinib. Residues with favorable and unfavorable energies have its values colored in green and red, respectively. The residues belonging to the ATP-binding site are highlighted in yellow and the residue that can bear a mutation in orange. Not all residues are presented in this table, only the ones that contribute with absolute values superior to a cut-off of 4 kcal/mol. This rule does not apply for the ATP-binding site residues, since we are interested in their contribution. Std= standard deviation. 115

Table 9: KIT residues contribution to the final binding energy of imatinib. Residues with favorable and unfavorable energies have its values colored in green and red, respectively. The residues belonging to the ATP-binding site are highlighted in yellow and the residues that can bear the mutation among the different systems, in orange. Not all residues are presented in this table, only the ones that contribute with absolute values superior to a cut-off of 4 kcal/mol. This rule does not apply for the ATP-binding site residues, or the mutation sites, since we are interested in their contribution. Residues containing an asterisk are present only at KIT^{V560G}. Std= standard deviation. 116

Table 10: Hydrogen bonds (H-bonds) occurrences between the targets and imatinib, averaged over the two MD replicas. The atom pairs for donor and acceptor interactions are depicted in the table. Imatinib atoms participating in the interaction are represented in the 2D representation of the inhibitor's structure. 130

Table A 1: Models issued from the Abinitio.relax application, selected by the distance between their N- and C-terminal residues. Models are ranked accordingly to their energy. Model ID can range from 1 to 10.000. The value in parentheses correspond to the deviation amount in relation to the specified distance value of 13 Å. 140

Table A 2: Best models generated by Modeller during the independent essays associated with each one of the six models selected from Rosetta. Each Modeller essay was composed of 100 runs and the Table is ranked by the best models from each essay, based on their DOPE score..... 141

Table A 3: Cluster analysis performed on the generated models from the loopmodel application. Structures were grouped into seven clusters, with the cluster ID ranging from O to 6. In the Table, the clusters are ranked by the score of the top seven best models, indicated by their Model IDs, that ranges from 1 to 431..... 142

Table A 4: Clustering data obtained from the CABS-fold run using consensus modeling default temperature parameters. The clusters are ranked according to their density, from the most populated to the least populated one. 143

List of figures

Figure 1: ***Structural organization of RTK III receptors.*** *Receptor tyrosine kinases of type III comprise an extracellular cytokine binding region subdivided into five domains (from D1 to D5), a single transmembrane (TM) helix, a juxtamembrane region (JMR), a conserved tyrosine kinase domain (TKD) composed of two lobes, separated by a kinase insert domain (KID), and a C-terminus tail. The letter P in orange circles represents the main phosphorylation sites implicated in receptor activation. The KID is represented as a dotted line since we do not know its tridimensional structure. 3*

Figure 2: ***Structure of RTK III cytoplasmic region.*** *Crystallographic structures of the native receptor CSF-1R in the inactive auto-inhibited (PDB ID: 2OGV) and the active forms (PDB ID: 3LCD) are taken for illustration and presented as cartoon. The different domains of CSF-1R and key structural fragments are highlighted in color. The N-terminal proximal lobe (N-lobe) is in blue, the C-terminal distal lobe (C-lobe) is in green, together with the pseudo-KID present at the inactive structure, the Cα-helix is in cyan, the activation loop (A-loop) is in red, the juxtamembrane region (JMR) is in orange. Represented as sticks are the DFG motif (Asp796, Phe797, Gly798) and a insertion showing the positive dipole created by the negative cap of the Asp residue at position 802 (816 in KIT) in the small 3-10 helix of the A-loop. The small helix is supposed to be stabilized by hydrogen bonds between the helix and the P-loop residues. 4*

Figure 3: ***Regulation of macrophage and osteoclast development by CSF-1****. Circulating CSF-1, produced by endothelial cells in blood vessels, together with locally produced CSF-1, regulates the survival, proliferation and differentiation of mononuclear phagocytes and osteoclasts. CSF-1 synergizes with hematopoietic growth factors (HGFs) to generate mononuclear progenitor cells from multipotent progenitors, and with receptor activator of NF-kB ligand (RANKL) to generate osteoclasts from mononuclear phagocytes. Red arrows indicate cell differentiation steps; blue arrows indicate cytokine regulation. Entirely description from (PIXLEY & STANLEY, 2004)..... 6*

Figure 4: ***CSF-1R structure highlighting the phosphorylation sites and putative downstream molecules that associate, via phosphotyrosine binding domains****. The tyrosine residues are numbered according to the mouse CSF-1R sequence with human sequence numbers in brackets and residues known to be phosphorylated in v-fms only in italics. Figure from (MOUCHEMORE & PIXLEY, 2012)..... 7*

Figure 5: ***Kinase inhibitor binding modes.*** *Kinase inhibitor–protein interactions are depicted by ribbon structures (left) and chemical structures (right). The chemical structures depict hydrophobic regions I and II of ABL1 (shaded beige and yellow respectively) and hydrogen bonds between the kinase inhibitor (inhibitor atoms engaged in hydrogen bonds to hinge are highlighted in green or to allosteric site in red) and ABL1 are indicated by dashed lines. The DFG motif (pink), hinge and the A-loop of ABL1 are indicated in the ribbon representations. The kinase inhibitors are shown in light blue.* ***a****. ABL1 in complex with the type 1 ATP-competitive inhibitor PD166326 (Protein Data Bank (PDB) ID 1OPK). Shown here is the DFG-in conformation of the A-loop (dark blue).* ***b****. The DFG-out conformation of the activation loop of ABL1 (dark blue) with the type 2 inhibitor imatinib (PDB ID 1IEP). The allosteric pocket exposed in the DFG-out conformation is indicated by the blue shaded area (right). Figure adapted from (ZHANG; YANG & GRAY, 2009). 13*

Figure 6: ***Proposed mechanisms of KIT activation by mutations****. The multi-states equilibrium of KIT cytoplasmic region in* KIT^{WT} *(A),* $KIT^{D816H/V/Y/N}$ *and* $KIT^{V560G/D}$ *(B: upper and lower panels). Each KIT conformation is represented as a molecular surface, except the JMR and the A-loop and imatinib drawn as cartoons and sticks respectively. In KIT mutants, the mutation position is shown by a ball. Equilibrium between two states is denoted by arrows of different thicknesses. (A) In the absence of SCF,* KIT^{WT} *is*

mainly in the inactive auto-inhibited state maintained by the JMR non-covalently bounded to the kinase domain. This state of KIT is the imatinib target. (B) Upper panel: The A-loop mutations (D816V/H/Y/N) induce the inactive non-auto-inhibited state of KIT evidenced by the JMR departure from the kinase domain. This effect conducts to deployment of the A-loop eventually leading to the constitutively active KIT state. The inactive non auto-inhibited state is not a suitable target for imatinib that inhibit the inactive auto-inhibited state. Lower panel: The JMR mutations (V560G/D) greatly impact the JMR binding to the kinase domain and facilitate its departure, favoring the non auto-inhibited state, whereas the inactive conformation of the A-loop is still conserved. The inactive non auto-inhibited state of KIT is more consented in KITV560G/D than in KITWT and especially in KITD816V/H/Y/N, leaded to the increased sensitivity of KITV560G/D to inhibitor compared to KITWT. In each panel, the most preferred state of KIT in the presence of imatinib is encircled. Figure and legend extracted from (CHAUVOT DE BEAUCHÊNE et al., 2014)..... 18

*Figure 7: **Sequence and structural alignment between CSF-1R and KIT**. Above: Structure alignment of CSF-1R (PDB ID: 2OGV) and KIT (PDB ID: 1T45) are represented in cartoon, colored in orange and wheat, respectively. Below: Sequence alignment of CSF-1R and KIT, taking as reference the sequences deposited in the Uniprot database: P07333, residues 539-972 for CSF-1R; P10721, residues 546-976 for KIT. Identical residues are colored in grey; similar residues are colored in black. The boxes colored in blue, green and red represent, respectively the JMR, hinge and A-loop residues. The residue mutated in the 802/816 position is highlighted in pink..... 19*

*Figure 8: **Schematic representation of the allosteric transition in hemoglobin**. (a) Ribbon diagram representation of tetrameric hemoglobin (PDB ID 1GZX). The proposed pathway responsible for the cooperative transition from tensed (T) to relaxed (R) is highlighted with red spheres and the heme groups are represented as light blue stick. (b) Allosteric transition of tetrameric hemoglobin, as proposed by Perutz (PERUTZ, 1970; PERUTZ et al., 1998). Tetrameric hemoglobin in the T state is depicted on the left with the two α-subunits (blue) and the two β-subunits (purple) each with their own heme group (light blue). Salt bridges, depicted as the red positive and blue negative charges, hold the molecule in the T conformation, and these salt bridges are released upon binding of oxygen (orange oval) in the transition to the R conformation (on the right) accompanied by a 15° turn of the subunits relative to each another. Also contributing to the equilibrium are 60 additional water molecules preferentially binding the R state. Extracted from (MOTLAGH et al., 2014) 21*

*Figure 9: **Interaction network and modular network representation in KIT cytoplasmic region**. Top: The interaction network between the A-loop tyrosine (Y823) and the catalytic loop residues (H790, D792 and N797) is depicted for WT KIT (A), the D816V mutant (MU) (B) and the D816V/D792E double mutant (C). The average conformation obtained from molecular dynamics is represented in pale cyan cartoons. H-bonds are displayed when their occupancy lies above 50% of the simulation time. Bottom: The modular network representations of the WT KIT (D), the D816V mutant (E) and the D816V/D792E double mutant (F) built by MONETA are depicted, focusing on the JMR, catalytic loop and A-loop regions. The average conformation obtained from molecular dynamics is represented in transparent cartoons. Communication pathways generated from residue in position 792 are displayed as chains of small black spheres connected together by black lines. The initial residue is highlighted by a bigger sphere centered on its Cα. The path linking the A-loop and the JMR through the catalytic loop is highlighted in magenta. Residues D/V816 and Y823 in the A-loop, D/E792 in the catalytic loop and V559 in the JMR are highlighted in licorice and labeled. Adapted from (LAINE; AUCLAIR & TCHERTANOV, 2012) 22*

*Figure 10: **Schematic representation of the energy potentials related to bonded and non-bonded interactions.** Left. (A) Representation for the bond potential, where l represents the bond length. (B)*

Representation for the angle potential to any three bonded atoms, where ϑ is the angle between three consecutive bonds. (C) Improper dihedral potential, where φimp is the angle between two planes. This potential is important to keep, for ex, the planarity of benzene rings. (D) Dihedral (proper) potential, where φ is representing the angle with torsion freedom. Right. Lennard-Jones graphic representation with atomic representations of the sum of the atomic vdW radius. Figure adapted from (FERNANDES, 2014) 27

*Figure 11: **Graphical representation for the first derivative minimization methods**. The green and the red lines at the 2D representation of the energy well correspond, respectively, to the steepest descent and the conjugate gradients methods. Source: Wikipedia 30*

*Figure 12: **Comparative modeling process**. The first step involves the search and selection of homologue structures to be aligned with the target sequence. The alignment will serve as a backbone in which the model will be constructed. The next step involves the adjustment of the alignment using data derived from secondary structure prediction, for ex. After the model construction, its validation is done through using softwares as Watcheck (HOOFT et al., 1996) and Procheck (LASKOWSKI et al., 1993), using the Ramachandran plot as criteria, for example. This graph permits to determine which torsion angles (ψ e ϕ) from the amino acid residues are correct, giving an idea about the precision of the model. Regions in red, brown and yellow represent, respectively, the favored, allowed and `generously allowed` defined by Procheck. Figure is reproduced from (BISHOP; DE BEER & JOUBERT, 2008). 37*

*Figure 13: **2D representation of the periodic boundary conditions in an infinite environment.** The central box is replicated on the three Cartesian coordinates and a cutoff radius, r_{cut}, can be used to restrict the region where the long-range interactions will be accounted for calculation during the MD simulation. Figure reproduced from http://wiki.cs.umt.edu/classes/cs477/index.php/Distance_Matrix#Periodic_boundary_conditions (accessed at 01/20/2015)..... 44*

*Figure 14: **Cutoff methods for treating the long-range interactions**. Represented in the graph are the adjustments by shift (red) or switch (green) of the interaction energy, $\varepsilon ij *$ (blue), between two atoms, i and j, in function of the interatomic separation, $Rij *$..... 45*

*Figure 15: **Schematic representation of the Modular MONETA's general principle**. A modular network representation composed of clusters of residues and chains of residues is built from the dynamical correlations and topology calculated from a protein conformational ensemble. In MONETA, residue clusters or modules are delineated as independent dynamic segments (IDSs) as they represent the most striking features of the protein local dynamics. Chains of individual residues are designated as communication pathways (CPs) as they represent well-defined connectivity pathways along which interactions can be mediated at long distances in the protein. Information is propagated through IDSs via the modification of the local atomic fluctuations and through CPs via well-defined interactions. The highly connected residues, at the junction of many pathways, can be considered as "hubs" in the protein network. Figure extracted from (ALLAIN et al., 2014)..... 54*

*Figure 16: **Overview of the major analysis steps in the MONETA workflow**. Each step of the MONETA procedure is delimited by an icon. The required inputs, parameters, outputs and scripts are identified by colors: initial mandatory inputs in purple, outputs in blue, MONETA computation steps in green, software and program in grey. Step 3 is illustrated by a 2D graph of the communication landscape in KIT (a) and by 3D representation of communication pathway in STAT5 (b). 2D and 3D graphs drawn with GEPHI and PyMOL modules incorporated in MONETA. Figure extracted from (ALLAIN et al., 2014) 56*

Figure 17: ***MD simulation data for CSF-1R inactive form****. Two forms of receptor, the native (CSF-1R^{WT} and CSF-1R^{MU} (D802V) were simulated twice during 50 ns. (**A**) The Root Mean Square Deviation (RMSD) values were calculated for backbone atoms from trajectories 1 and 2 of MD simulations of CSF-1R^{WT} (black and blue) and CSF-1R ^{MU} (red and orange). RMSDs (in nm) plotted versus simulation time (ns) and showed separately for N- and C-lobes, JMR and A-loop regions. (**B**) The Root Mean Square Fluctuations (RMSF) computed on the backbone atoms over the total production simulation time of CSF-1R^{MU} (red) were compared to those in CSF-1R^{WT} (black). The RMSFs of the A-loop is zoomed in the insert. The average conformations for CSF-1R^{WT} (**C**) and CSF-1R^{MU} (**D**) are presented as tubes. The size of the tube is proportional to the by-residue atomic fluctuations computed on the backbone atoms. The high fluctuation region found in proteins, are specified by red color and numerated from 1 to 10 in B–D. The size of numbers in **D** is proportional to RMSFs. (DA SILVA FIGUEIREDO CELESTINO GOMES et al., 2014) 78*

Figure 18: ***MD conformations of CSF-1R cytoplasmic region in the native protein and its D802V mutant****. Ribbon diagrams display the proteins regions or fragments with different colors: JMR (orange), A-loop (red), N- and C-lobe (blue and green), and KID (gray). Snapshots taken from the two MD replicas at 15, 25, 35 and 50 ns for CSF-1R^{WT} (top) and CSF-1R^{MU} (bottom) were superimposed by pair. Superposed conformations were selected by RMSDs clustering. The change in the A-loop conformation in CSF-1R^{MU} is highlighted with a black box. 80*

Figure 19: ***Secondary structure prevalence during MD runs.*** *Secondary structure assignments for JMR (A) and A-loop (B) were averaged over the two 50-ns MD simulations of CSF-1R^{WT} and CSF-1R^{MU}. For each residue, the proportion of every secondary structure type is given as a percentage of the total simulation time. Each secondary structure type is shown with lines of different colors: 3_{10} helices (in cyan), parallel β-sheet (in red), turn (in orange), bend (in blue) and bridge (green). Coiled structure is shown by dashed purple lines. The D802V position is indicated as a red circle. 81*

Figure 20: ***Secondary structure prediction of the JMR sequence (residues 538–580) from CSF-1R^{WT}****. The prediction was performed using sequence-based algorithms GOR4, Jpred, SOPMA, SCRATCH, NetSurfP, Psipred and a structure-based method STRIDE. Predicted structural elements are coded as indicated at bottom.(DA SILVA FIGUEIREDO CELESTINO GOMES et al., 2014) 83*

Figure 21: ***Distance monitoring between the JMR and the TK domain of CSF-1R****. Left: Distances d1 and d2 between the centroids C1 (JM-B)) and C1' (N-lobe) and between C2 (JM-S) and C2' (C-lobe), respectively. Right : Distance d1 (at the top) and d2 (at the bottom) monitored during the two replicas of the 50 ns MD simulations (full and dashed lines) for CSF-1R^{WT} (black) and CSF-1R^{MU} (red). 84*

Figure 22: ***Principal component analysis (PCA) of CSF-1R cytoplasmic region in the inactive state.*** *The calculation was performed on the backbone atoms of CSF-1R^{WT}and CSF-1R^{MU}. Top: (**A**) The bar plot gives the eigenvalues spectra of CSF-1R^{WT}(blue) and CSF-1R^{MU} (orange) in descending order. (**B**) The grid gives the overlap between the first 10 eigenvectors from CSF-1R^{WT} (columns) and CSF-1R^{MU} (rows). The overlap between two eigenvectors is evaluated as their scalar product and represented by colored rectangles, from blue (0) through green and yellow to red (0.51). Bottom: Modes 2 and 3 atomic components for CSF-1R^{WT} (**C**) and CSF-1R^{MU} (**D**) are drawn as yellow arrows on the protein cartoon representation. JMR is in blue, A-loop is in violet and the rest of protein is in grey. (DA SILVA FIGUEIREDO CELESTINO GOMES et al., 2014) 85*

Figure 23: ***Convergence analysis of the MD simulations for CSF-1R^{WT} (WT) and CSF-1R^{MU} (D802V) models performed on the 90 ns concatenated trajectories.*** *Grouping of MD conformations was done using five independent runs calculated for each model. The populations of each group for each run are presented as histograms in the logarithmic scale denoted by different colors, black and grey from the*

1st and 2nd halves of the two replicas, respectively. The identification numbers of each reference structure denotes the time (ns) in which it was picked from the MD trajectory. The fourth run of A and B contains reference structures that are better represented in both replicas and it was chosen for further NM calculations..... 87

*Figure 24: **Binding energy changes between CSF-1R^{WT}and CSF-1R^{MU} in the inactive state.** Left : A thermodynamic cycle picturing the dissociation of JMR from KD in CSF-1R^{WT} (blue) and CSF-1R^{MU}(red). Right: The total free energy (ΔG) of the JMR binding to the kinase domain, computed over the individual MD simulations for both CSF-1R^{WT}and CSF-1R^{MU}. (DA SILVA FIGUEIREDO CELESTINO GOMES et al., 2014) 89*

*Figure 25: **H-bond patterns in CSF-1R stabilizing the auto-inhibited inactive state of CSF-1R^{WT}and the non-inhibited inactive state of CSF-1R^{MU}.** H-bonds between residues from (A) JMR and Cα-helix; (B)JMR and C- loop and (C) A-loop and C-loop. Snapshots taken from the most representative conformations derived from MD simulations by the convergence analysis. All residues presented as sticks, in blue for CSF-1R^{WT}and in orange for CSF-1R^{MU}. The H-bonds are shown as dotted lines, red and green in CSF-1R^{WT}and CSF-1R^{MU} respectively. (D) The DFG motif conformation together with JMR's anchoring residue W550. Representation of DFG and W550 residues conformations originated from the crystallographic structure (2OGV, green) and representative MD conformations of CSF-1R^{WT} (blue) and CSF-1R^{MU} (orange). 90*

*Figure 26: **Comparison of the JMR sequence in CSF-1R and KIT and Electrostatic Potential (EP) surface in the two receptors.** (**A**) The amino acids composition of JMR (JM-B and JM-S) in CSF-1R and KIT. (**B**) The EP surface of TK domain and JMR in two receptors, CSF-1R and KIT. EP calculations on the Connolly solvent-accessible surfaces of the receptors were performed with the APBS software. The color scale ranges from red (electronegative potential) through white (neutral) to blue (electropositive potential). (DA SILVA FIGUEIREDO CELESTINO GOMES et al., 2014) 92*

*Figure 27: **Independent dynamic segments and communication pathways in cytoplasmic region of CSF-1R.** Top: Structural mapping of the Independent Dynamic Segments (IDSs) identified in CSF-1R^{WT} (**A**) and CSF-1R^{MU} (**B**). The average conformations are presented as tubes. IDSs were identified from the analysis of the merged 60 ns concatenated trajectory. IDSs are referred to as S_i, where i = 1, 2,…,N, labeled and specified by color in the both proteins. The largely modified or newly found IDSs in the mutant are referred to as S'_i in red. Bottom: 3D structural mapping of the inter-residues communication in CSF-1R^{WT} (**C**) and CSF-1R^{MU} (**D**), computed over the last 30 ns of the individual MD simulations. MD 2 is taken for illustration. The average MD conformation is presented as cartoon. The proteins fragments are presented with different colors: JMR (blue), Cα-helix (cyan), P-loop (yellow), C-loop (green) and A-loop (red). Communication pathways (CPs) between residues atoms (small circles) are depicted by colored lines: CPs formed by the A-loop residues are represented in orange; by the JMR-residues in magenta. The key residues in the communication networks are labelled and depicted as bulky circles. (DA SILVA FIGUEIREDO CELESTINO GOMES et al., 2014). 94*

*Figure 28: **3D structural mapping of the inter-residues communication in CSF-1R^{WT}, CSF-1R^{MU}, KIT^{WT} and KIT^{MU}**. The average MD conformation is presented as cartoon. The proteins fragments are presented with different colors: JMR (blue), Cα-helix (violet), P-loop (yellow), C-loop (green) and A-loop (red). Communication pathways (CPs) between residues atoms (small circles) are depicted by coloured lines: CPs formed by the A-loop residues in orange; by the JMR-residues in magenta. The key residues in the communication networks are labelled (in WT) and depicted as bulky circles. (DA SILVA FIGUEIREDO CELESTINO GOMES et al., 2014) 99*

Figure 29: ***Structures of the cytoplasmic domain of CSF-1R and KIT in the native form****. Superimposition of the CSF-1R and KIT crystallographic structures: (**A**) CSF-1R (2OGV) and KIT (1T45) in the inactive conformation; (**B**) CSF-1R in the inactive (2OGV) and the active conformations (3LCD) ; (**C**) KIT in the inactive (1T45) and active (1PKG) conformations. The proteins are presented as cartoon, CSF-1R is in blue light and KIT is in grey light. The key structural fragments of receptors in the inactive and the active conformations are highlighted in color. The JMR is in yellow and in orange; the A-loop is in red and magenta; the Cα-helix is in cyan and blue. The relative orientation of the Cα-helix (inserts) in the two proteins is presented together with the principal axis of helices detected with PyMol. (DA SILVA FIGUEIREDO CELESTINO GOMES et al., 2014)..... 100*

Figure 30: ***2D representation of the chemical structure of imatinib****. The labeled atoms in the figure constitute the ligand's atoms that are engaged in H-bonds interactions with the protein ATP-binding site residues. Hydrogen in N7 represents the protonation site..... 101*

Figure 31: ***Imatinib bound to KIT in its inactive form (PDB ID: 1T46)****. Structures of auto-inhibited CSF-1R (PDB ID: 2OGV) and KIT (PDB ID: 1T45) were superimposed to highlight the side chain orientations of the residues Trp located at the JMR and the DFG-motif Phe. Imatinib is represented in sticks colored in pink, residues Trp and Phe from CSF-1R (550) and KIT (811) are colored in blue and green, respectively. In orange is represented the Phe from KIT when complexed with imatinib (1T46), the Trp is located on a missing part of the JMR in this crystal. 102*

Figure 32: ***Selected conformations for the docking simulations****. Superimposition of the ATP-binding site residues, described at (MOL et al., 2004) with their corresponding in the structures of CSF-1R (A) and KIT (B) selected by the convergence analysis and used in the docking simulations. Residues are represented in sticks and labeled following the numbering of CSF-1R and KIT separately. Imatinib and the ATP-binding site residues of crystal 1T46 are represented as sticks and colored in magenta. (A) Residues corresponding to CSF-1R^{WT} and CSF-1R^{D802V} are colored in pale green and wheat, respectively. (B) Residues corresponding to KIT^{WT}, KIT^{V560G}, KIT^{S628N} and KIT^{D816V} are colored in cyan, orange, green and black, respectively. 103*

Figure 33: ***Validation of the docking methodology.*** *Superimposition of the docking best poses for WT CSF-1R (pale green) and KIT (cyan) with the crystallographic structure of WT KIT complexed with imatinib (magenta) referred by the PDB code 1T46 (MOL et al., 2004). Imatinib and the surrounding ATP-binding residues are represented as sticks and the labels correspond to CSF-1R and KIT numbering, respectively. In the crystal, imatinib makes H-bonds with T670, C673, E640 and D810; H790 and I789 are described as hydrophobic contacts. 104*

Figure 34: ***Best docking poses for each CSF-1R and KIT systems****. Imatinib is represented in orange sticks and the protein backbone is represented in grey as cartoon with the residues that interact with Imatinib in the crystal structure 1T46 are depicted represented as grey sticks and depicted in the first frame. Hydrogen bonds between the protein and the ligand are represented as dotted lines..... 105*

Figure 35: ***Comparison between the binding-site conformations pre- and after docking simulations****. ATP-binding site residues described in (MOL et al., 2004) are represented as sticks and colored as magenta (before docking) and green (after docking). Imatinib is represented as orange sticks. 106*

Figure 36: ***RMSD for the protein backbone atoms, excluding the C-ter tail.*** *RMSD values calculated for CSF-1R (A) and KIT (B) complexes in their WT apo, WT complexed with imatinib and mutant forms complexed with imatinib. The initial protein coordinates before the MD simulations were used as reference. Curves corresponding to replicas 1 and 2 are colored in blue and orange, respectively. (C) Superposition of the crystallographic structures of CSF-1R (2OGV) and KIT (1T45) as grey cartoon with key TK domain elements represented in different colors: JMR in orange and yellow, Cα-helix in blue and*

cyan, A-loop in pink and red, for CSF-1R and KIT, respectively. ATP-binding site where imatinib is placed is represented by an ellipse. 108

*Figure 37: **RMSF for protein backbone atoms**. RMSF values for CSF-1R (above) and KIT (below). The different forms of the receptor are designated in the legend. 109*

*Figure 38: **RMSF 3D representation on the protein backbone, excluding the KID and the C-ter regions**. The different conformations of CSF-1R and KIT are labeled. The regions that fluctuate the most are thicker, colored in red and numerated in the apo receptor forms. 111*

*Figure 39: **Binding energy between imatinib and the WT and mutant forms of CSF-1R and KIT**. Below the graph, the experimental IC_{50} values for the inhibition are indicated. The letters S and R indicate if the analyzed form of the receptor is sensitive or resistant to imatinib, respectively. The superscript + indicate more sensitivity or more resistance. The asterisk on the R associated with S628N mutation was placed because we do not know with certitude the resistance character of the mutation. 113*

*Figure 40: **Binding energy decomposition**. The main components that contribute to the final binding energy, according to the MMPBSA approach, are represented. 114*

*Figure 41: **Difference between CSF-1R^{WT} and CSF-1R^{D802V} residue contribution to the binding energy**. Residues presenting differences superior to 1 or inferior to -1 kcal/mol are highlighted. 119*

*Figure 42: **Difference between KIT^{WT} and each mutant for the residue contribution to the binding energy**. Residues presenting differences superior to 1 or inferior to -1 kcal/mol are highlighted. The energy contribution for the truncated portion of the JMR was not considered. 119*

*Figure 43: **Electrostatic surface profile for all the receptor structures**. The color code corresponding to the charged nature of the surface is place in the bottom. For illustration, imatinib is placed in the ATP-binding sites and represented as sticks colored in cyan. The view of the ATP-binding site corresponds to the cavity where the protonated nitrogen is situated. 121*

*Figure 44: **Salt-bridges profile calculated over the MD simulation replicas for WT CSF-1R and the mutant**. The occurrence of the salt bridges are represented in percentage of the MD simulations time. 122*

*Figure 45: **Salt-bridges profile calculated over the MD simulation replicas for WT KIT and the mutants**. The occurrence of the salt bridges are represented in percentage of the MD simulations time 123*

*Figure 46: **Binding energy between imatinib and the WT and mutant forms of KIT.** Data correspond to the new simulations test data in which all KIT forms contain the same truncated portion of the JMR, as in the complex formed by the mutant KIT^{V560G}. 125*

*Figure 47: **RMSD for the backbone protein atoms**. The JMR and the C-ter tail were excluded from the calculation. All the systems are labeled accordingly to each KIT form; KIT^{V560G} previous data are presented for comparison reasons. Replicas 1 and 2 are colored in blue and orange, respectively... 126*

*Figure 48: **RMSF calculated for protein backbone atoms**. The different forms of KIT are label as indicated in the legend. 127*

*Figure 49: **RMSF 3D representation on the protein backbone, excluding the JMR, KID and the C-ter regions**. The different conformations of KIT are labeled. The regions that fluctuate the most are thicker, colored in red and numbered in the structure of the WT. 127*

*Figure 50: **RMSF for the residues situated in a radius of 6 Å around imatinib**. RMSF values were calculated for the backbone and each bar correspond to each form of KIT. Residues that are engaged*

in H-bonds interactions with imatinib accordingly to the crystal structure 1T46 are highlighted in bold. 128

Figure 51: ***2D representation of imatinib and the ATP-binding site residues involving in H-bond interactions with the inhibitor.*** *(A) H-bonds pattern concerning the systems KIT^{WT} and KIT^{V560G}. We see that the side chain orientation of the protein residues remains intact. (B). H-bonds concerning the systems KIT^{S628N} and KIT^{D816V}.In both systems, the side chain orientation of Asp810 changes, which facilitates the H-bond between AspOδ and N7 from imatinib. The bonds coloured in green and orange represent the contacts that were lost in the second run of MD simulations, for KIT^{S628N} and KIT^{D816V}, respectively..... 131*

Figure A 1: ***Primary sequence of the kinase insert domain (KID)****. In black, is represented the real sequence of the KID and in green and red, respectively, are the extra residues used in the modeling protocols and the secondary structure analysis..... 144*

Figure A 2: ***Blast most significant results.*** *The crystal structure of CSF-1R 3LCO is the one that has the best coverage for the KID sequence, containing the residues 680-686, 747-751. ..... 145*

Figure A 3: ***Comparison between the sequences for the KID region among all the members of type III RTK family****. Conserved residues (*), conservative mutations (:) and semi-conservative mutations (.) are colored in red, green and blue, respectively..... 145*

Figure A 4: ***Secondary structure prediction for CSF-1R's KID using different methods****. The consensus prediction was done manually; regions without a consensus are represented by a hyphen. The burial residue index was generated by SAM_T08 program. 146*

Figure A 5: ***Prediction of the disorder tendency for the CSF-1R's KID residues calculated by the IUPRED web-server.*** *Values above 0.5 indicate disordered structures (DOSZTÁNYI et al., 2005). 146*

Figure A 6: ***Final models generated de novo by Rosetta****. The 10.000 models generated by the Abinitio.relax module were sorted by energy and the 100 lowest energy structures were analyzed using a distance criteria of 13 Å length between the N- and C- terminals. Only six models corresponded to the distance criteria. In this figure, the models are numerated accordingly to their energy score from the lowest to the highest energy model. Structures are colored by their secondary structure: α-helices in red, β-sheets in yellow and coil in green. For comparison, the consensus secondary structure prediction is represented using the same color code from the models..... 147*

Figure A 7: ***Final models for the CSF-1Rf^{ull} generated using the Protocol 1****. The models are numbered in the figure accordingly to figure A 6. The TK domain region modeled from the template 2OGV is represented in blue and the KID region predicted by Rosetta is colored by their secondary structure as in figure A 6. ..... 148*

Figure A 8: ***Best models outputted by Rosetta following the Protocol 2****. The TK domain region modeled using Modeller is colored in blue and the KID predicted region is colored by their secondary structure as in figures 31 and 32. The regions denoted by – in the sequences are already modelled, using the PDB 3LCO as template. 150*

Figure A 9: ***Best models outputted by the CABS-fold web-server following the Protocol 2****. The TK domain region modeled using Modeller is colored in blue and the KID region predicted by CABS is colored by their secondary structure as in figures 31 and 32. The regions denoted by – in the sequences are already modelled, using the PDB 3LCO as template..... 151*

Abbreviations and acronyms

ABL	Abelson murine leukemia viral oncogene homolog
A-loop	Activation loop
AMBER	Package to Perform MD simulations
Amber99sb	Amber forcefield, parameter set 99sb
AML	Acute myeloid leukemia
APBS	Adaptive Poisson-Boltzmann Solver
CFF	Classical Force Fields
ASICs	Application-specific instruction chips
CHARMM	Package to perform MD simulations
Charmm27	CHARMM force field with parameters for all atoms
C-helix	α-helix C from N-terminal lobe of the tyrosine kinase receptors
C-lobe	C-terminal lobe
C-loop	Catalytic loop
CLUSTAWL	Multiple sequence alignment server
CML	Chronic myelomonocytic leukemia
CPs	Communication pathways
CPUs	Computing processing units
CSF-1	Colony stimulating factor-1
CSF-1R	Colony stimulating factor-1 receptor
$CSF\text{-}1R^{apo}$	apo form of CSF-1R, in absence of ligand
$CSF\text{-}1R^{full}$	Full length CSF-1R cytoplasmic domain
$CSF\text{-}1R^{WT}$	wild-type form of CSF-1R
$CSF\text{-}1R^{MU}$/$CSF\text{-}1R^{D802V}$	mutant form of CSF-1R bearing the D802V mutation
CT	Commute time
DFG	Sequence motif located in the A-loop, composed of three amino-acids: Asp, Phe and Gly.
DIMB	Diagonalization in a mixed basis method
DSSP	Database of secondary structure assignment
EGF	Epidermal growth factor
EM	Energy minimization
ENMs	Elastic network models
EP	Electrostatic potential
FGF	Fibroblast growth factor
FLT3	FMS-like tyrosine kinase 3
GISTs	Gastro-intestinal stromal tumors
Gleevec	Commercial name for imatinib, Novartis
GB	Generalized Born
GOR	secondary structure prediction based on information theory
GPUs	Graphical processing units
GROMACS	Package to perform MD simulations
H-bond	Hydrogen bond
HGFs	Hematopoietic growth factors
HMM	Hidden Markov model

IDSs	Independent dynamic segments
IFD	Induced fit docking
JM-B	Juxtamembrane binder region
JM-S	Juxtamembrane switch region
JM-Z	Juxtamembrane zipper region
JMR	Juxtamembrane domain
JPRED	Secondary prediction method based on PSSM and HMMs
KD	Kinase domain
KIC	Kinematic closure
KID	Kinase insert domain
KIT	Stem cell growth factor receptor
KIT^{apo}	Apo form of KIT, in absence of ligand
KIT^{D816V}	mutant form of KIT bearing the D816V mutation
KIT^{S628N}	mutant form of KIT bearing the S628N mutation
KIT^{WT}	wild-type form of KIT
LFA	Local feature analysis
MD	Molecular dynamics
MM-GBSA	Molecular Mechanics Generalized Born Surface Area
MM-PBSA	Molecular Mechanics Poisson-Boltzmann Surface Area
MONETA	Modular network analysis
NetSurfP	Secondary structure prediction based on neural networks
NGF	Nerve growth factor
N-lobe	N-terminal lobe
NMA	Normal modes analysis
NPT	Isothermal-isobaric ensemble
NVE	Microcanonical ensemble
NVT	Canonical ensemble
ns	nanoseconds
PB	Poisson-Boltzmann
PBC	Periodic boundary conditions
PCA	Principal component analysis
PDB	Protein data bank
PDGFR-α	Platelet-derived growth factor alpha
PDGFR-β	Platelet-derived growth factor beta
PME	Particle Mesh Ewald
PSSM	Position-specific scoring matrix
PSIPRED	Secondary structure prediction server based on PSSM
PTB	Phosphotyrosine-binding domain
PyMOL	Molecular visualization program
QM	Quantum Mechanics
RANKL	Receptor activator of NF-kβ ligand
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
RTKs	Receptor tyrosine kinases
SCF	Stem cell factor
SCFR	Stem cell factor receptor

SCRATCH	Secondary structure predictor based on a combination of different methods
SH2	Src homology 2 domain
SOPMA	Secondary structure prediction method based on multiple-alignments with known secondary structure sequences
STI571	prototype denomination for imatinib
STRIDE	Knowledge based secondary structure predictor
TAMs	Tumor-associated macrophages
TK	Tyrosine kinase
TKIs	Tyrosine kinase inhibitors
TM	Transmembrane domain
VDW	van der Waals
VEGF	Vascular endothelial growth factor
WT	Wild-type

Resumo

Os receptores tirosino-quinase (RTKs) CSF-1R (*colony stimulating fator-1 receptor*) e SCFR ou KIT (*stem cell fator receptor*) são importantes mediadores da sinalização celular associada à proliferação, sobrevivência e diferenciação de macrófagos e células da linhagem hematopoiética, respectivamente. A função basal desses receptores é alterada por mutações "ganho-de-função" que induzem sua ativação constitutiva seguida da sinalização celular descontrolada, associada a vários tipos de câncer e doenças inflamatórias. Essas mutações também alteram a sensibilidade dos receptores aos inibidores de TKs, como o imatinib, empregado na quimioterapia contra diferentes tipos de câncer. A mutação V560G, localizada no domínio justamembranar (JMR) do receptor KIT aumenta a sensibilidade do receptor ao imatinib, enquanto que as mutações S628N e D816V em KIT e a mutação D802V em CSF-1R induzem a resistência ao medicamento, sendo as duas últimas localizadas do loop de ativação do receptor (A-loop). O JMR e o A-loop constituem dois segmentos regulatórios que sofrem grande mudança conformacional durante a ativação dos receptores, pela perda de interação entre seus resíduos e o restante da proteína. Os objetivos dessa tese são (i) investigar os efeitos estruturais e dinâmicos induzidos pela mutação D802V na porção intracelular do CSF-1R, comparando os resultados com os obtidos para a forma nativa do CSF-1R, assim como os dados obtidos anteriormente para a mutação D816V em KIT; (ii) caracterizar a afinidade do imatinib às formas nativas e mutantes do domínio TK dos receptores KIT (V560G, S628N e D816V) e CSF-1R (D802V), correlacionando as predições computacionais com os dados experimentais disponíveis na literatura. Por meio de simulações de dinâmica molecular (DM), mostramos que as mutações D802V, em CSF-1R, e D816V, em KIT, tem efeitos diferentes. A mutação D802V tem um impacto local na estrutura do A-loop, além de interromper a comunicação alostérica entre este segmento e o domínio JMR. Contudo, a ruptura desses caminhos de comunicação não é suficiente para induzir o destacamento do JMR em relação ao domínio TK, devido à presença de ligações hidrogênio bastante prevalentes durante o tempo de simulação. O efeito mais sutil da mutação em CSF-1R também foi associado à diferença na sequencia primaria de ambos receptores na sua forma nativa, principalmente na região do JMR. Isto poderia explicar porque essa mutação é pouco observada no câncer. Na etapa seguinte, caracterizamos a afinidade do imatinib aos diferentes alvos através de simulações de ancoramento e DM, além do cálculo da energia livre de ligação. Os dados de energia se mostraram coerentes com os dados experimentais de inibição para as formas selvagens e mutantes dos receptores. A decomposição da energia nos diferentes termos que contribuem para a ligação mostrou o termo eletrostático como o principal determinante da diferença de energia entre os tipos selvagens e mutantes resistentes. As mutações D802V e D816V mostraram-se as mais deletérias na contribuição para a ligação do imatinib, devido à redistribuição de cargas positivas ao redor do sitio de fixação do ligante. Nossos dados também sugerem que o JMR tem um papel minoritário no mecanismo de resistência.

Abstract

The receptors tyrosine kinase (RTKs) for the colony stimulating factor-1 (CSF-1R) and the stem cell factor (SCFR or KIT) are important mediators of signal transduction related to the proliferation, survival and differentiation of macrophages and cells from the hematopoietic lineage, respectively. The normal function of these receptors can be compromised by gain-of-function mutations that lead to the constitutive activation of the receptors, associated with cancer diseases and inflammatory disorders. A secondary effect of the mutations is the alteration of the receptor's sensitivity to tyrosine kinase inhibitors, such as imatinib, compromising the use of these molecules in the clinical treatment. The mutation V560G in the juxtamembrane (JMR) domain of KIT increases the receptor's sensitivity to imatinib, while the mutations S628N and D816V, in KIT, and D802V in CSF-1R, trigger resistance. The last two being located at the activation loop (A-loop) of the receptors. The JMR and the A-loop constitute two regulatory fragments that undergo a dramatic conformational change during the receptors' activation, due to the loss of essential interactions with the rest of the protein. Our goals in this thesis consisted in (i) study the structural and dynamics effects on the intracellular domain of CSF-1R induced by D802V mutation and compare the results with those obtained for KIT in the native wild-type (WT) and mutated forms; (ii) study the affinity of imatinib to the WT and mutant forms of the TK domain of KIT (V560G, S628N and D816V) and CSF-1R (D802V), correlating the computational predictions with the available experimental data. By means of molecular dynamics (MD) simulations, we have showed that the D802V mutation in CSF-1R does not produce the same dynamic and structural effects caused by the D816V mutation in KIT. The D802V mutation has a local impact on the A-loop structure and disrupts the allosteric communication between this fragment and the JMR. However, the disruption is not sufficient to induce the JMR's departure from the TK domain, due to the strong coupling between the JMR's distal region and the TK domain, stabilized by highly prevalent H-bonds. The subtle effect of the mutation in CSF-1R was associated with the difference in the primary sequence between both receptors in the native form, particularly in the JMR region, and this could explain why this mutation is not frequently found in cancer. In the following step, we have characterized by docking, MD simulations and energy calculations, the binding affinity of imatinib to the different targets. The free energy associated with the binding of imatinib was consistent with the experimental data. The energy decomposition in the different terms contributing to the binding energy evidenced that the electrostatic interactions are the main force that drives the sensitivity or the resistance of the targets to imatinib. The mutations D802V and D816V showed to be the most deleterious in the energy contribution to the binding of imatinib, due to the charge redistribution of positive charges in the vicinity of the binding site. Our data also indicated that the JMR domain has a minor role in the resistance mechanism.

Résumé

Les récepteurs à activité tyrosine kinase (RTKs) CSF-1R (*colony stimulating factor-1 receptor*) et KIT (*stem cell factor receptor*) sont médiateurs importants de la signalisation cellulaire associé à la prolifération, survie et différentiation des macrophages et cellules du lignage hématopoïétique, respectivement. La fonction basale des RTKs est altérée par des mutations « gain de fonction » qui induisent leur activation constitutive et une signalisation cellulaire tronquée associée à divers types de cancer et de maladies inflammatoires. Ces mutations modifient également la sensibilité des récepteurs aux inhibiteurs de TKs, comme l'imatinib, utilisé en clinique dans le traitement de différents cancers. Le mutant V560G, localisé sur le domaine juxtamembranaire (JMR) de KIT, augmente la sensibilité du récepteur à l'imatinib. En revanche, les mutations S628N et D816V dans KIT, et la mutation D802V dans CSF-1R induisent un phénomène de résistance. Ces deux dernières sont localisées sur la boucle d'activation (boucle A) des récepteurs. Le JMR et la boucle A constituent d'importants éléments régulateurs ; ils subissent un important changement de conformation lors de l'activation des RTKs, du à la perte d'interactions essentielles avec le reste de la protéine. Dans cette thèse, nos objectifs sont (i) d'étudier les effets structuraux et dynamiques induits par la mutation D802V chez la partie intracellulaire de CSF-1R et les comparer aux résultats obtenus pour la forme sauvage de CSF-1R et également les données concernant la mutation D816V chez KIT ; (ii) caractériser l'affinité de l'imatinib aux formes sauvages et mutantes du domaine TK de KIT (V560G, S628N et D816V) et CSF-1R (D802V), corrélant les prédictions computationnelles avec les données expérimentales. Par simulations de Dynamique Moléculaire (DM), nous avons établi que la mutation D802V chez CSF-1R n'entraine pas les mêmes effets structuraux provoqués par la mutation D816V chez KIT. La mutation D802V provoque un effet locaux sur la structure de la boucle A et interrompt la communication allostérique entre ce fragment et le JMR. Néanmoins, la rupture de ces chemins de communication n'est pas suffisante pour induire le départ du JMR, des liaisons hydrogène assurant le couplage du JMR au domaine TK pendant les simulations de DM. L'effet subtil de la mutation chez CSF-1R a également été attribué aux différences de séquence primaire des deux récepteurs dans leur forme sauvage, surtout dans la région du JMR. Ceci pourrait expliquer pourquoi cette mutation est peu observée dans le cancer. Dans l'étape suivante, nous avons caractérisé l'affinité de l'imatinib aux différentes cibles par simulations de *docking*, DM et calculs d'énergie libre. L'énergie de liaison de l'imatinib aux formes sauvages et mutées de KIT et CSF-1R est corrélée aux données expérimentales. La décomposition des différents termes de la valeur finale d'énergie libre a montré que les interactions électrostatiques constituent la force motrice de la sensibilité ou de la résistance. La substitution de l'aspartate dans la position 802/816 pour une valine est la plus délétère en termes d'énergie, en raison d'une redistribution de charges positives á proximité du site de fixation de l'imatinib. Nos données ont également indiqué que le JMR a un rôle mineur dans le mécanisme de résistance.

Chapter 1: Introduction

1. Tyrosine kinase receptors

Receptor tyrosine kinases (RTKs) are cell-surface transmembrane receptors that possess a tightly regulated tyrosine kinase (TK) activity within their cytoplasmic domain (BLUME-JENSEN & HUNTER, 2001). They act as sensors for extracellular ligands, the binding of which triggers receptor dimerization and activation of the kinase function, leading to the recruitment, phosphorylation and activation of multiple downstream signaling proteins, which ultimately govern the physiology of cells (HUBBARD & MILLER, 2007).

RTKs family includes the receptors for insulin and for many growth factors, such as epidermal growth factor (EGF), fibroblast growth factor (FGF), platelet-derived growth factor (PDGF), vascular endothelial growth factor (VEGF), and nerve growth factor (NGF) (HUBBARD & TILL, 2000). Because of their role as growth factor receptors, many RTKs have been implicated in the onset or progression of various cancers, either through gain-of-function mutations or through receptor/ligand overexpression (BLUME-JENSEN & HUNTER, 2001).

RTKs structure consist of an extracellular portion that binds polypeptide ligands, a transmembrane helix and a cytoplasmic portion that possesses tyrosine catalytic activity. Most of RTKs proteins are monomeric in ligand absence and the extracellular portion typically contains a discrete array of globular domains such as immunoglobulin (Ig)-like domains, fibronectin type III–like domains, cysteine-rich domains, and EGF-like domains. In contrast, the domain composition of the cytoplasmic portion is simpler, consisting of a juxtamembrane region (JMR), followed by the tyrosine kinase (TK) catalytic domain and a carboxy-terminal region (Fig. 1). Some receptors, most notably members of the PDGF receptor family, contain a large insertion of ~100 residues in the TK domain, known as the kinase insert domain (KID).

The specific reaction catalyzed by tyrosine kinase proteins is the transfer of the phosphate cleaved from ATP to the hydroxyl group of a tyrosine in a protein substrate. Activation of RTKs requires generally two steps: enhancement of intrinsic catalytic activity and creation of binding sites to recruit downstream signaling proteins (HUBBARD & TILL, 2000). Both of these processes are achieved through autophosphorylation on tyrosine residues, which is a

consequence of ligand-induced oligomerization, which induces a structural rearrangement of the receptor intracellular portion, facilitating the tyrosine autophosphorylation.

In general, autophosphorylation of tyrosines in the activation loop (A-loop) results in kinase activity and autophosphorylation of tyrosines in the JMR, KID and carboxy-terminal regions generates binding sites for modular domains of the cellular signaling proteins or modulators that recognize phosphotyrosine specific sequences. The two well-established phosphotyrosine binding modules present within signaling proteins are the Src homology 2 (SH2) domain and the phosphotyrosine-binding (PTB) domain (KURIYAN & COWBURN, 1997).

Receptor autophosphorylation can occur in *cis* (within a receptor) or in *trans* (between receptors). Structural studies on the insulin receptor kinase domain indicate that the A-loop tyrosines in TKs can only be phosphorylated in *trans*. Other phosphorylation sites (e.g. JMR or carboxy-terminal tail) could potentially be autophosphorylated in *cis*.

2. Type III RTK subfamily

Based on their overall architecture and kinase domain (KD) sequence, RTKs have been grouped into 20 subfamilies (ROBINSON; WU & LIN, 2000). Also known as the PDGF receptor family, type III RTK subfamily includes the stem cell factor (SCF) receptor KIT, the macrophage colony-stimulating factor-1 (CSF-1) receptor CSF-1R (or FMS), the platelet-derived growth factor α and β (PDGFR-α and PDGFR-β) and the FMS-like tyrosine kinase 3 (FLT3) (ROBINSON; WU & LIN, 2000; ULLRICH & SCHLESSINGER, 1990). In this work, we have studied RTKs that belong to type III RTK subfamily of receptors, with special interest in the CSF-1R and KIT.

Type III RTKs share the common RTK structure, being their extracellular domain composed of Ig-like units, followed by a single-pass transmembrane helix, the JMR coupled to the cytoplasmic TK domain including a KID (HONEGGER, 1990; ROSNET & BIRNBAUM, 1993) of a variable length (~ 60-100 residues), and a carboxy-terminal tail (GOLDBERG *et al.*, 1990; ROSNET *et al.*, 1993; VERSTRAETE & SAVVIDES, 2012) (Fig. 1).

The TK domain has a bi-lobar structure, with an ATP-binding cleft located between the N- and C-terminal lobes. The N-lobe is composed of twisted five-stranded anti-parallel β-sheet adjacent to an α-helix (Cα-helix) and the C-lobe shows predominantly α-helical structure (Fig.

2). The C-lobe contains an activation loop (A-loop) that begins with the highly conserved 'DFG' motif composed of three amino acids – aspartic acid (D), phenylalanine (F), and glycine (G).

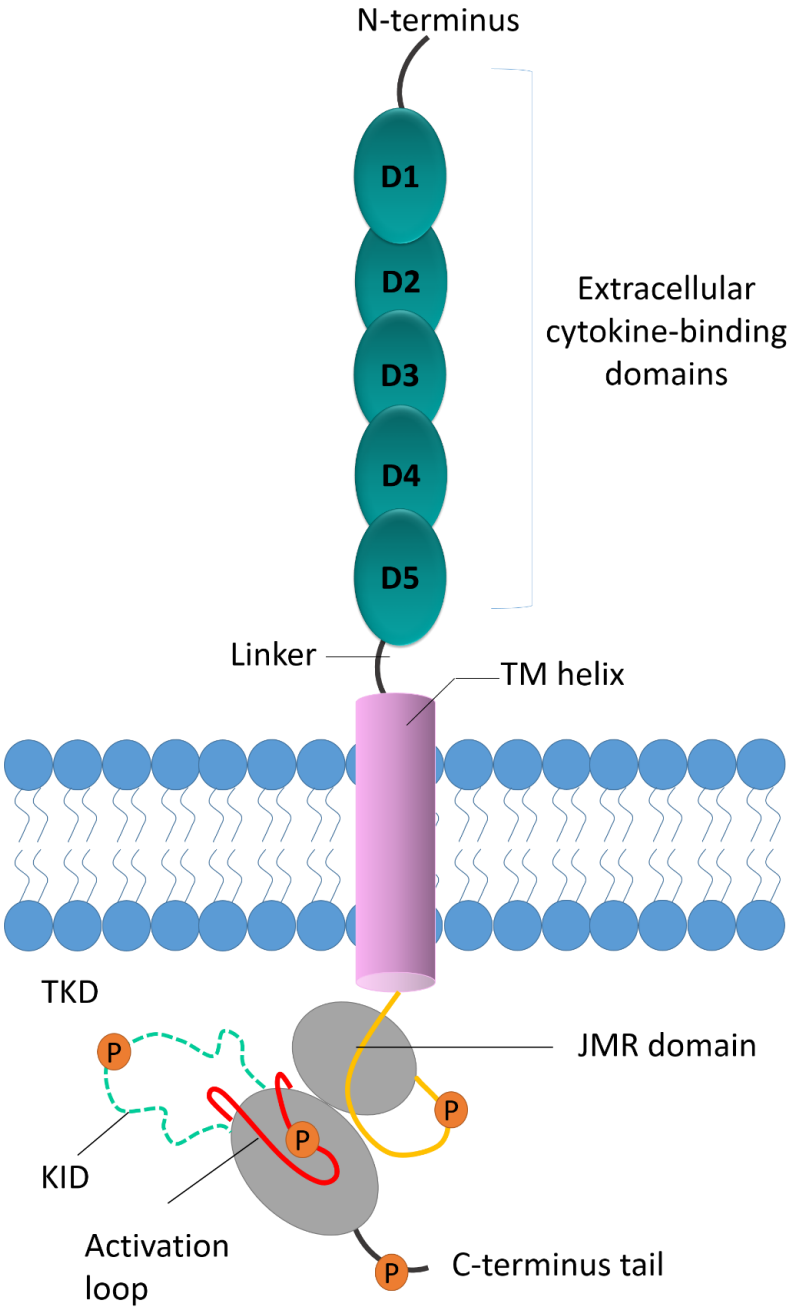


Figure 1: ***Structural organization of RTK III receptors.*** *Receptor tyrosine kinases of type III comprise an extracellular cytokine binding region subdivided into five domains (from D1 to D5), a single transmembrane (TM) helix, a juxtamembrane region (JMR), a conserved tyrosine kinase domain (TKD) composed of two lobes, separated by a kinase insert domain (KID), and a C-terminus tail. The letter P in orange circles represents the main phosphorylation sites implicated in receptor activation. The KID is represented as a dotted line since we do not know its tridimensional structure.*

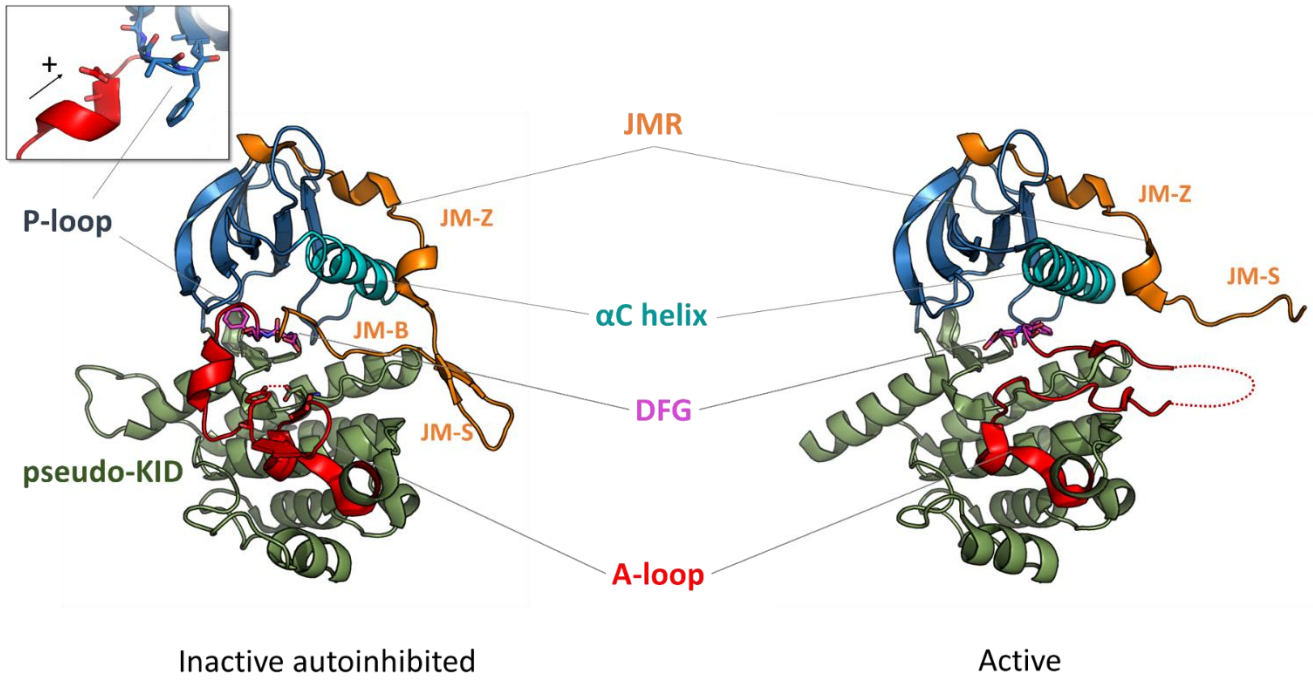


Figure 2: ***Structure of RTK III cytoplasmic region.*** *Crystallographic structures of the native receptor CSF-1R in the inactive auto-inhibited (PDB ID: 2OGV) and the active forms (PDB ID: 3LCD) are taken for illustration and presented as cartoon. The different domains of CSF-1R and key structural fragments are highlighted in color. The N-terminal proximal lobe (N-lobe) is in blue, the C-terminal distal lobe (C-lobe) is in green, together with the pseudo-KID present at the inactive structure, the Cα-helix is in cyan, the activation loop (A-loop) is in red, the juxtamembrane region (JMR) is in orange. Represented as sticks are the DFG motif (Asp796, Phe797, Gly798) and a insertion showing the positive dipole created by the negative cap of the Asp residue at position 802 (816 in KIT) in the small 3-10 helix of the A-loop. The small helix is supposed to be stabilized by hydrogen bonds between the helix and the P-loop residues.*

In the absence of ligand, the receptors are in dynamic equilibrium between two states: the inactive autoinhibited state that is highly dominant, and the active state (LANDAU & BEN-TAL, 2008; WAN & COVENEY, 2011). Two crucial kinase regulatory segments, the A-loop and the JMR, undergo extensive conformational rearrangements during the activation/deactivation processes (Fig. 2).

Several intramolecular interactions contribute to keep the receptor in the auto-inhibited inactive form. In the inactive state of the receptor, the A-loop is adjacent to the active site and the DFG motif, in its N-extremity, adopts an “out” conformation, *i.e.,* its phenylalanine is flipped into the ATP-binding site, thus preventing ATP and Mg^{2+} co-factor binding (GRIFFITH *et al.*, 2004; MOL *et al.*, 2004). This conformation is stabilized by the JMR that inserts itself directly into the kinase active site and impairs the arrangement of the A-loop in its active conformation. The single tyrosine in the A-loop binds to the catalytic loop as a pseudo-substrate and contributes to keep the receptor in its inactive form. Upon activation, the JMR

moves from its auto-inhibitory position to a completely solvent-exposed emplacement. This is followed by a conformational swap of the A-loop from its inactive packed arrangement to an active extended conformation. Such large-scale conformational transition, together with a switch of the DFG motif to an “in” conformation allows ATP entrance and binding in the catalytic site.

Analysis of the crystallographic structures of KIT, CSF-1R and FLT3 in their inactive state (GRIFFITH *et al.*, 2004; MOL *et al.*, 2004; WALTER *et al.*, 2007) suggested that the JMR has also an important role in the mechanism of auto-inhibition, based on extensive interactions with the TK domain. The JMR is composed of three fragments: JM-Binder (JM-B), buried into the TK domain making direct contacts with the Cα-helix, the catalytic (C-) loop and the A-loop; JM-Switch (JM-S) that adopts a hairpin-like conformation positioned apart from the C-lobe and contains the tyrosine residues responsible for the conformational switch; and JM-Zipper (JM-Z), packed along the solvent-exposed face of the Cα-helix (Fig. 2). Together, the JM-B and the JM-Z block the Cα-helix, which also regulates the catalytic activity of the kinases (LI *et al.*, 2003), and prevent the A-loop from adopting an active conformation, restricting the inter-lobe plasticity.

2.1. CSF-1 and its receptor CSF-1R

CSF-1 is the most pleiotropic macrophage growth factor, stimulating the survival, proliferation and differentiation of mononuclear phagocytes. Most tissue macrophages and osteoclasts are regulated by CSF-1 (Fig. 3). The effects are mediated through autophosphorylation of its receptor, CSF-1R, and the subsequent phosphorylation of downstream molecules. (PIXLEY & STANLEY, 2004).

CSF-1 is a homodimeric growth factor that is expressed in different isoforms (secreted glycoprotein/proteoglycan or cell surface isoform) with different locals of actions and effects. While CSF-1R is the sole receptor for CSF-1, an alternative functional ligand for the receptor, interleukin-34 (IL-34) was recently identified (LIN *et al.*, 2008). IL-34, a dimeric glycoprotein broadly expressed in different tissues, supports monocyte and macrophage survival and proliferation. Despite these similarities, the two cytokines show little or no primary homology at a peptide level, although a structural modeling study suggests that IL-34 folds in a four-helix

bundle structure (GARCEAU *et al.*, 2010), similar to CSF-1, what would justify the interaction in the receptor. Nevertheless, it was revealed a different spatiotemporal expression pattern, suggesting different and independent roles for both cytokines (WEI *et al.*, 2010).

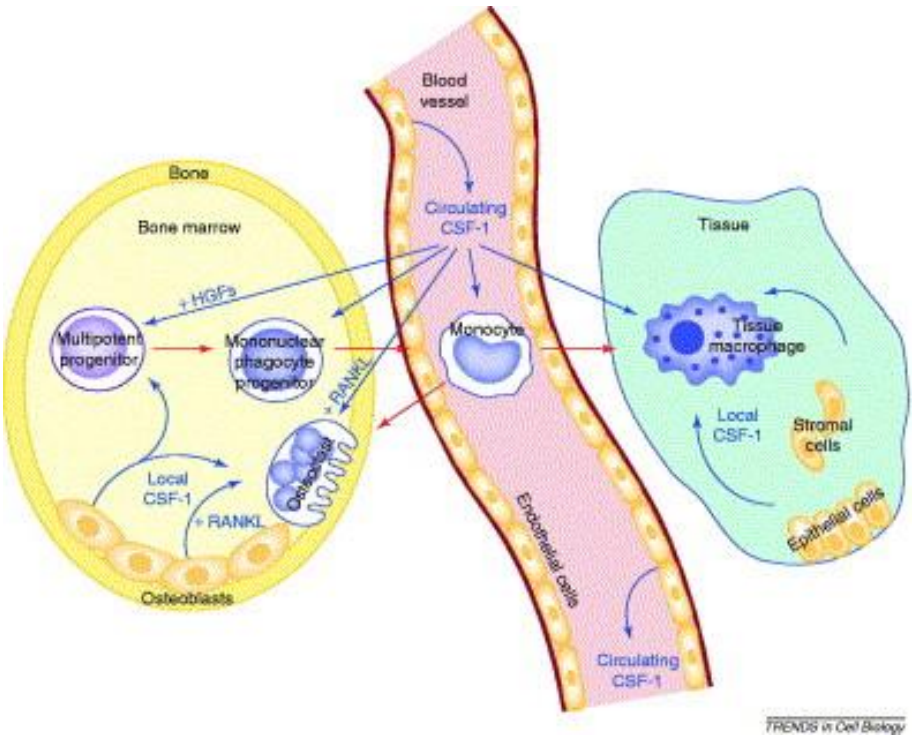


Figure 3: ***Regulation of macrophage and osteoclast development by CSF-1****. Circulating CSF-1, produced by endothelial cells in blood vessels, together with locally produced CSF-1, regulates the survival, proliferation and differentiation of mononuclear phagocytes and osteoclasts. CSF-1 synergizes with hematopoietic growth factors (HGFs) to generate mononuclear progenitor cells from multipotent progenitors, and with receptor activator of NF-kB ligand (RANKL) to generate osteoclasts from mononuclear phagocytes. Red arrows indicate cell differentiation steps; blue arrows indicate cytokine regulation. Entirely description from* (PIXLEY & STANLEY, 2004)

2.1.1. Cell signaling pathways activated by CSF-1R

CSF-1R is a member of type III RTKs subfamily and its general activation mechanism was already discussed in the precedent sections. The receptor is activated when homodimeric CSF-1 binds to D1-3 of its extracellular domain (Fig. 4), inducing receptor dimerization, initially non covalent but bridged by a disuphide bond (LEMMON & SCHLESSINGER, 2010). Of the 19 tyrosine residues in the intracellular domain of the CSF-1R, six have been shown to be phosphorylated in response to CSF-1, Y559, Y697, Y706, Y721, Y807 and Y974 (numbered

according to murine CSF-1R sequence), with phosphorylation of Y544 and Y921 demonstrated only in the constitutively active v-fms oncoprotein (JOOS *et al.*, 1996) (Fig. 4).

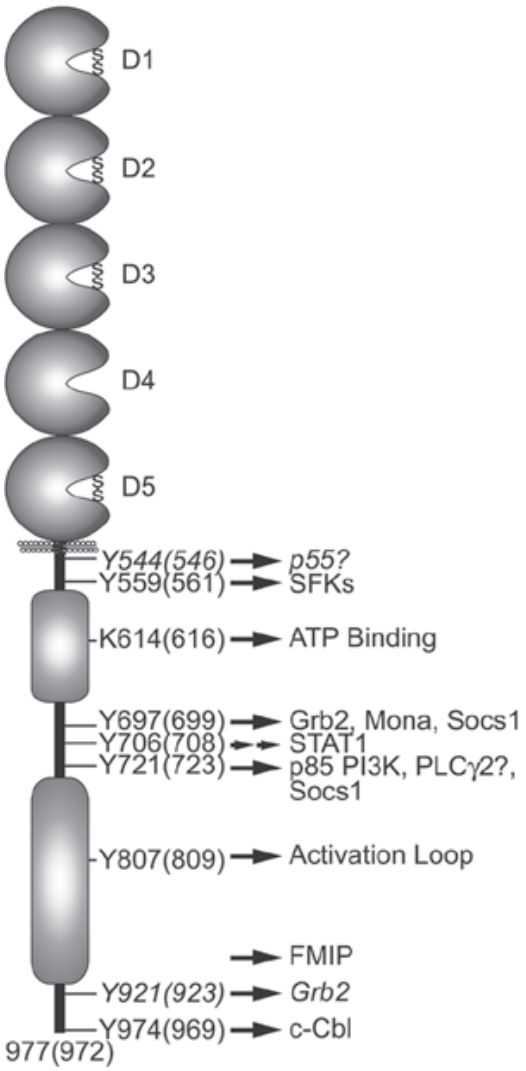


Figure 4: ***CSF-1R structure highlighting the phosphorylation sites and putative downstream molecules that associate, via phosphotyrosine binding domains****. The tyrosine residues are numbered according to the mouse CSF-1R sequence with human sequence numbers in brackets and residues known to be phosphorylated in v-fms only in italics. Figure from* (MOUCHEMORE & PIXLEY, 2012)

However, recent studies have revealed that Y544F CSF-1R exhibits markedly reduced *in vitro* kinase activity and *in vivo* tyrosine phosphorylation, including Y559 phosphorylation (YU *et al.*, 2008, 2012). Y544 (Y546 in human CSF-1R) is highly conserved through type III RTK subfamily and it is located in the a buried region of the JMR, making key hydrogen bonds with a key conserved residue that regulates interlobe plasticity, E633, located at the Cα-helix (WALTER *et al.*, 2007).

Phosphorylation of the majority of these CSF-1R tyrosine residues creates specific binding motifs for CSF-1-induced association of phosphotyrosine binding domain-containing effector molecules that are themselves substrates of the receptor (MOUCHEMORE & PIXLEY, 2012). Domains that mediate the binding of the effectors with receptor phosphotyrosine motifs include SH2 and PTB domains (SCHLESSINGER & LEMMON, 2003). Once the partner proteins are activated, it is initiated a series of phosphorylation cascades that leads to cytoskeletal remodeling and increased adhesion, as well as increased transcription and translation necessary for the pleiotropic effects of CSF-1.

The role of individual tyrosine residues in signaling pathways involving CSF-1R is reviewed by (MOUCHEMORE & PIXLEY, 2012). Briefly discussing that, among the known pathways implicated in macrophage survival, differentiation, proliferation and motility, two major pathways can be highlighted. The CSF-1 induced cytoskeletal remodeling are regulated through PI3K by association with Py721 and production of PIP_3. The Rho family of GTPases are also implicated in this pathway. PI3K-stimulated PIP_3 production also produces translocation and activation of the serine/threonine kinase, Akt, to trigger a number of downstream effectors involved in cell survival, proliferation and motility.

The second major pathway activated by CSF-1 is the Ras/Raf/MEK/ERK or MAPK pathway, which is deregulated in many types of cancer. Like the PI3K/Akt pathway, the MAPK pathway regulates many fundamental cellular processes. This pathway is activated following phosphorylation of Y697 and Y921, which induces translocation of Grb2 to the receptor. Grb2 initiates a series of signaling events that end up in the activation of ERK1 and ERK2, the main effectors of this cascade. ERK1/2 together phosphorylate over 70 known substrates, including several transcription factors that rapidly and transiently induce transcription of the immediate early response genes, c-fos, c-jun and c-myc. Induction of these proto-oncogenes stimulates DNA and protein synthesis to permit transition of macrophages through G1 phase of the cell cycle and into cell division.

2.1.2. The CSF-1 role in cancer

The relation of CSF-1 with cancer was first identified in breast, ovarian and endometrial cancers when high levels expression of the growth factor in this cells was correlated with poor outcomes of the disease (MCDERMOTT *et al.*, 2002; SCHOLL *et al.*, 1994). High tissue

expression levels of CSF-1 and its receptor correlate with breast and colon cancer metastasis to draining lymph nodes and with metastatic prostate cancer (RICHARDSEN *et al.*, 2008; WEBSTER *et al.*, 2010). As well as paracrine involvement of CSF-1/CSF-1R signaling in cancer spread, CSF-1 and CSF-1R co-expression by tumor cells produces autocrine stimulation of tumorigenesis in breast, ovarian and endometrial cancer (PATSIALOU *et al.*, 2009).

Despite the implication of wild-type CSF-1R in all this events necessary for tumor progression and metastasis, there is only a few evidences for activating CSF-1R mutations in malignance tumors, in the contrary of other members of type III RTK subfamily. As compared to KIT, whose activating mutations are hallmarks of systemic mastocytosis (PARDANANI, 2013), and gastro-intestinal stromal tumors (GISTs) (CORLESS *et al.*, 2005), or to FLT3, whose activating mutations are frequently observed in acute myeloid leukemias (AML) (SWORDS; FREEMAN & GILES, 2012), activating mutations in CSF-1R gene have been rarely detected in human tumors (SOARES *et al.*, 2009).

Preliminary observations of leukemogenic point mutations in CSF-1R at L301 and Y969 were not confirmed later (RIDGE *et al.*, 1990; SUCH *et al.*, 2009). These two mutations would be located at the extracellular portion and KID, respectively, two regions difficult to study *in silico* since there is no available structural data.

Nevertheless, CSF-1R is a therapeutic target in oncology, either to inhibit the paracrine loop that promotes tumor growth or to re-educate tumor-associated macrophages (TAMs) within tumor microenvironment (PYONTECK *et al.*, 2013). CSF-1 and CSF-1R could also be targeted in non-cancer diseases. CSF-1R high levels expression are found in inflammation processes, which are known to contribute to cancer development. Some of the inflammation processes include rheumatoid arthritis (BISCHOF *et al.*, 2000; YANG *et al.*, 2006), lupus nephritis (MENKE *et al.*, 2009) and atherosclerosis (CLINTON *et al.*, 1992).

2.2. KIT receptor and hotspot mutations

KIT is the receptor for the stem cell factor (SCF), which governs the proliferation and differentiation of the hematopoietic cells (ZSEBO *et al.*, 1990). KIT signaling is also implicated in the activation and degranulation of mastocytes, which induces the liberation of inflammation agents (COLUMBO *et al.*, 1992). In addition, KIT is expressed by embryonic

melanoblasts (MCCULLOCH & MINDEN, 1993) and melanocytes from derma and epidermis in adults (GRICHNIK *et al.*, 1998) and by germ cells in male gonads (MCCULLOCH & MINDEN, 1993).

RTKs oncogenic mutations are supposed to alter the equilibrium between the inactive and active forms of the receptor by disrupting the auto-inhibitory interactions that keep the receptor in the inactive form. The receptor becomes constitutively active, promoting signaling independently of ligand binding in the extracellular portion and induce exacerbated cell proliferation. KIT mutations are usually found at the extracellular domain, in the JMR and in the A-loop. Most of the mutations were observed to be implicated in gastrointestinal tumors (GISTs) and mastocytosis (abnormal accumulation of mastocytes in the skin, bone-marrow, digestive tube, etc); in about 17% of sinonasal T cell lymphomas; 9% of ovarian and testicle tumors and in 1% of acute myeloid leukemia (AML) cases (BOUGHERARA *et al.*, 2009).

GIST tumors related with KIT abnormal activity result mainly from deletions or substitutions located at the JMR, in codons 550 to 561, such as V560G/D (FLETCHER & RUBIN, 2007) and V559G mutations (DEBIEC-RYCHTER *et al.*, 2006) (2/3 of clinical cases). In a smaller proportion, are found deletions or insertions in the extracellular domain (~10%), substitutions on the nucleotide binding domain (K642E or V654A, 1 to 3%), and in the A-loop (1 to 3%) (GOUNDER & MAKI, 2011).

Located in the A-loop, D816V substitution in KIT is the most important mutation implicated in the mastocytosis (PRICE; GREEN & KIRSNER, 2010). Other substitutions were also identified in cases of systemic mastocytosis (D816Y/H/F) and AML (D816H/Y) (ASHMAN *et al.*, 2000; BEGHINI *et al.*, 1998; NING; LI & ARCECI, 2001).

Crystallographic structures of KIT are available for the wild-type (WT) in apo and ligand-complexed forms. Two structures, in particular, were solved complexed with sunitinib, in its WT and mutated form containing the D816H substitution (PDB IDs: 3G0E and 3G0F, respectively) (GAJIWALA *et al.*, 2009). The comparison between both structures has revealed that the mutation provokes a local effect on the structure of the A-loop helix located between residues 817-819. In the mutant, this region was not well folded, and this loss of structure could disturb the inactive conformation of the A-loop. In addition, the mutant shows that the N-terminal region of the JMR is displaced in respect to its position on the WT structure of KIT.

All these changes could indicate that the conformation presented at the crystal 3G0F correspond to an inactive non auto-inhibited form of KIT trapped by the ligand.

The authors (GAJIWALA *et al.*, 2009) have also reported that the crystallization of D816H mutant was only possible due to the depletion of part of the JMR's N-terminal (structure starts from residue 562 in this crystal). Deuterium-hydrogen exchange experiments confirmed the flexible nature of the JMR. Moreover, in the KIT mutant (mutation D816H), the JMR was more solvent-accessible. These results suggest that there is a long-range effect of the mutation D816H in the conformation and JMR behavior, and this effect could be extrapolated to D816V substitution.

In vitro experiments with KIT D816V/H and V560D mutants of KIT in absence of SCF showed that the mutants activate much faster than KIT WT and the mutant V560D is even more fast to activate than D816V/H mutants (GAJIWALA *et al.*, 2009). Cross-linking experiments suggested that the activation of the receptor KIT would be different in these two mutants: in $mM^{-1}s^{-1}$, WT (0.25) < D816H (46) < D816V (134) < V560D (>150). The signaling cascade of the mutants was also analyzed *in vivo*. These studies have revealed that cells expressing the D816V mutant have an abnormal recruitment and activation of Pi3K by STAT5, which would play a key role in the tumor genesis (HARIR *et al.*, 2008). In addition, the mutants present the constitutive activation of STAT3 via the Src and Fes kinases, which is not observed at the WT KIT, even in the presence of SCF (ZHAO & IYENGAR, 2012). All this results strongly confirm that the mutations does not only trigger the constitutive activation of the receptor but also they modify the signaling of the activated receptor.

A new mutation on KIT has been recently identified. A gain-of-function point mutation was found in a patient with metastatic melanoma. The mutation was located at exon 13 on the DNA obtained from a bone metastasis and confirmed in a lung metastasis from a biopsy. The substitution was a missense at codon 628, resulting in a S628N substitution (VITA *et al.*, 2014).

2.3. RTKs as targets in cancer chemotherapy

Advances in molecular biology in the past few years have provided some remarkable advances in the understanding of tumor biology and oncogenesis and in the targeted cancer therapy (KERR, 2003). As discussed early, RTKs dysfunction results from WT super expression

or gain-of-function mutations that lead to the constitutive receptor activation. Therefore, inhibiting the activation of RTKs is important and the receptors have become targets to many compounds targeting cancer diseases related to abnormal function of these receptors.

Kinases have become one of the most pursued classes of drug targets in the treatment of cancer and other disorders, such as immunological, neurological, metabolic and infectious disease. Some factors contribute to the popularity of kinases as drug targets: virtually every signal transduction process is wired through a phosphotransfer cascade, so inhibition would produce a real physiological response; despite a high degree of conservation in the ATP binding site, highly selective small molecules with favorable pharmaceutical properties can be developed; inhibition of kinase activity in normal cells can often be tolerated, presenting a therapeutic window for the selective killing of tumor cells (ZHANG; YANG & GRAY, 2009). However, despite these motivating factors, the field faces significant challenges, such as drug resistance, lacking of inhibitor selectivity and efficacy, for example.

As already mentioned, the protein kinases are defined by their ability to catalyze the transfer of the terminal phosphate of ATP to substrates. The ATP binding site is located on a cleft situated between the N- and the C-lobes of the TK domain. The adenine ring of the ATP forms H-bonds with the kinase “hinge” (the segment that connects the N- and the C- lobes) (Fig. 5) and the ribose and triphosphate groups of ATP bind in a hydrophilic channel extending to the substrate binding site that features conserved residues important to the catalysis. The substrate binding site is also known as the catalytic loop (C-loop).

The first developed tyrosine kinase inhibitors (TKIs) were ATP-competitive and classified as type I inhibitors. This type of molecules mimics the purine ring of the ATP adenine moiety and targets the active conformation of the kinase. They consist typically of a heterocyclic ring-based system that occupies the purine binding site, where it serves as a scaffold for side chains that occupy the adjacent hydrophobic regions I and II (Fig. 5).

Interestingly, the first inhibitor to reach the market was a compound that recognizes the inactive form of the kinase. Imatinib (Gleevec, Novartis) binds to the protein in the ‘DFG-out’ conformation, preventing the binding of both nucleotides and protein substrates. Imatinib inhibits the ABL1, KIT and PDGFR receptors and also CSF-1R (ZHANG; YANG & GRAY, 2009). The discovery of the binding mode of imatinib was followed by a new generation of inhibitors,

called type II. They bind in the same region occupied by type I inhibitors but extend to the additional hydrophobic site available in the inactive form (Fig. 5) (ZUCCOTTO *et al.*, 2010).

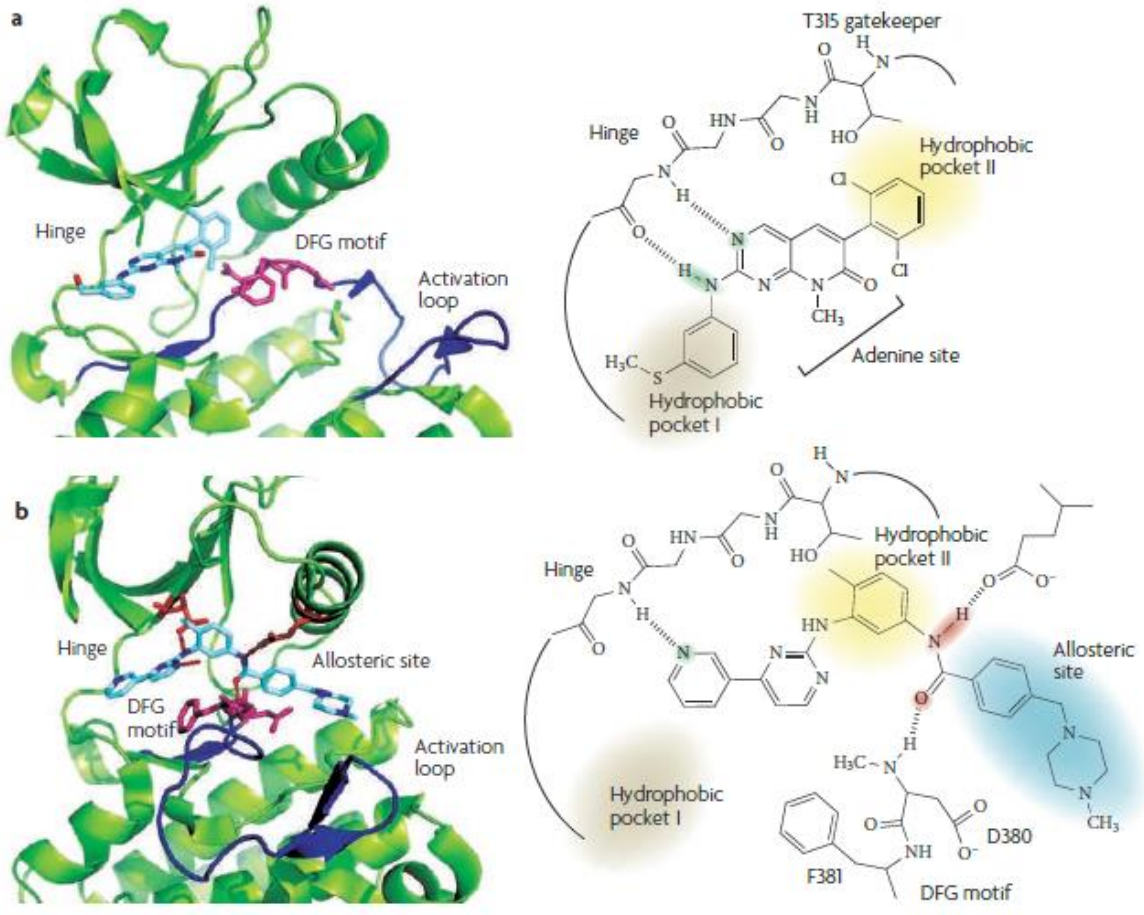


Figure 5: ***Kinase inhibitor binding modes.*** *Kinase inhibitor–protein interactions are depicted by ribbon structures (left) and chemical structures (right). The chemical structures depict hydrophobic regions I and II of ABL1 (shaded beige and yellow respectively) and hydrogen bonds between the kinase inhibitor (inhibitor atoms engaged in hydrogen bonds to hinge are highlighted in green or to allosteric site in red) and ABL1 are indicated by dashed lines. The DFG motif (pink), hinge and the A-loop of ABL1 are indicated in the ribbon representations. The kinase inhibitors are shown in light blue.* ***a****. ABL1 in complex with the type 1 ATP-competitive inhibitor PD166326 (Protein Data Bank (PDB) ID 1OPK). Shown here is the DFG-in conformation of the A-loop (dark blue).* ***b****. The DFG-out conformation of the activation loop of ABL1 (dark blue) with the type 2 inhibitor imatinib (PDB ID 1IEP). The allosteric pocket exposed in the DFG-out conformation is indicated by the blue shaded area (right). Figure adapted from* (ZHANG; YANG & GRAY, 2009).

The inhibitor-stabilized conformational rearrangement observed in structures presenting the bound inhibitor type-2 in kinase shows that the kinase active site can remodel to accommodate a variety of inhibitors (LIU & GRAY, 2006). For instance, structure of WT KIT complexed with Imatinib (PDB ID:1T46) revealed that the inhibitor displaces the auto-inhibitory domain from its position observed in the WT apo form of KIT (MOL *et al.*, 2004).

A third kind of inhibitors would be the *allosteric* compounds. Since they do not bind at the ATP-binding site, they modulate the kinase activity in an allosteric manner. These kinds of compounds exhibit the highest degree of kinase selectivity since they exploit binding sites that are unique to each particular kinase. The fourth class of kinase inhibitors are the covalent inhibitors. They form an irreversible covalent bond to the kinase active site by reacting with a nucleophilic cysteine residue (ZHANG; YANG & GRAY, 2009).

Mutational hotspots have been identified in type III RTKs and besides inducing tyrosine kinase constitutive activation, as discussed earlier for KIT and CSF-1R, these mutations can also alter the sensibility of the receptors towards TKIs. The TKI imatinib, highly specific for a restricted number of kinases, inhibits some of KIT mutants associated with GIST, in particular those carrying somatic mutations in the JMR. Mutants bearing the V560G/D substitution, in the JMR, are particularly sensitive, even more than the WT, to imatinib (FROST *et al.*, 2002). While the inhibitor is poorly efficient in the treatment of mastocytosis where the A-loop mutations D816V (H/Y/N) are present (CORLESS *et al.*, 2005; FROST *et al.*, 2002; HAYASHI *et al.*, 2001).

Acquired resistance to systemic therapy is still a challenge to the cancer treatment. KIT secondary mutations has been identified in naïve GIST patients treated with imatinib (DEMETRI *et al.*, 2002). These mutations are placed in the ATP-binding pocket or in the A-loop and lead to resistance to imatinib treatment (HEINRICH *et al.*, 2006). The second-line treatment, after imatinib failure, is provived by less specific inhibitors, such as sunitinib, that shows potency against imatinib-resistant KIT mutated in the ATP-binding pocket (V654A and T670I) (TAMBORINI *et al.*, 2004). Nevertheless, A-loop mutations such as D816V/H are equally resistant to second-line treatment inhibitors (FROST *et al.*, 2002).

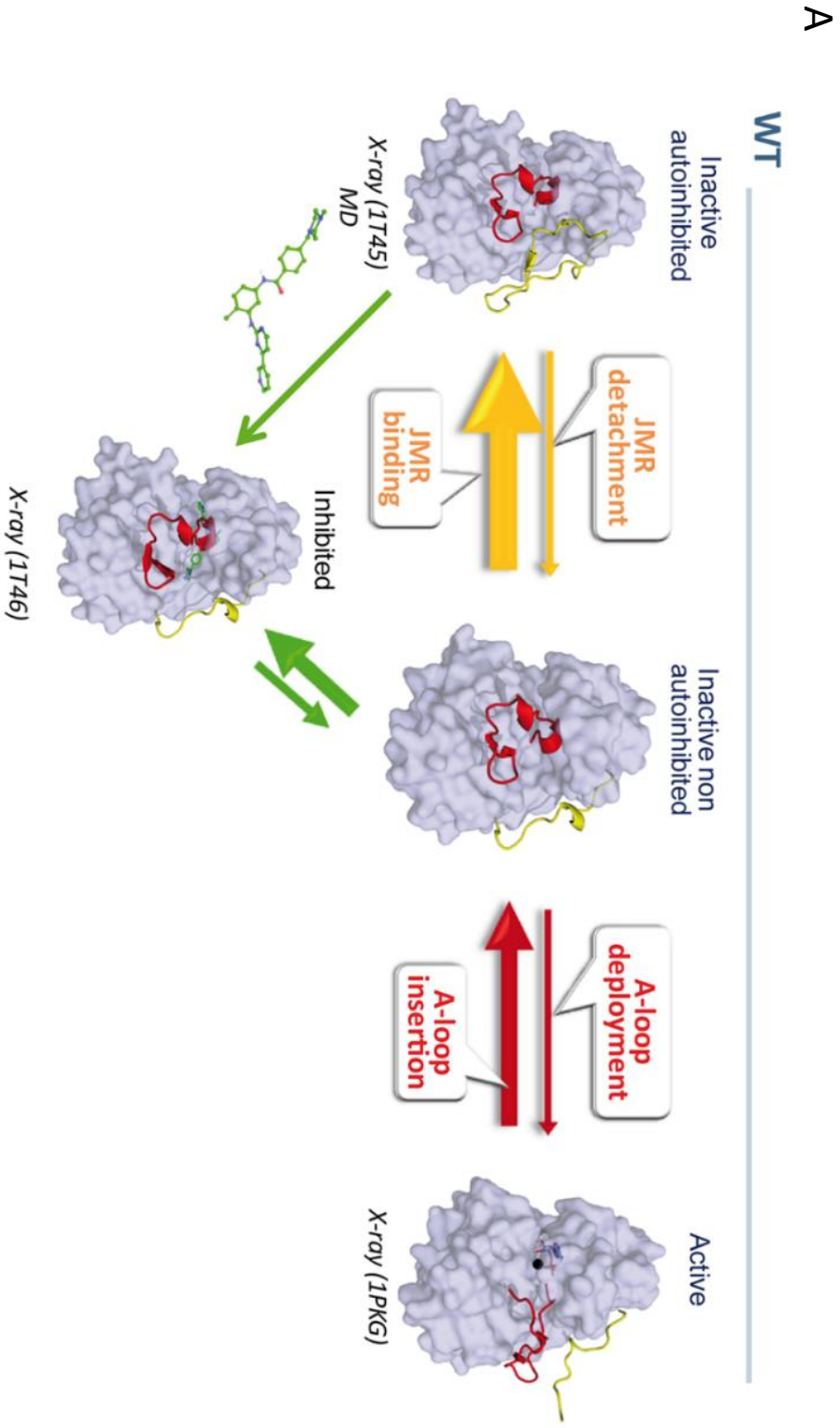
The role of S628N substitution is still not clear, although the authors declare that the mutant receptor is sensitive to imatinib, the inhibition of the autophosphorylation of the KIT S628N receptor occurred at a moderate concentration (1µM) of imatinib and dasatinib (also a second-line treatment inhibitor), when compared to the resistant KIT D816V receptor, used as positive control (IC_{50} > 5µM) (VITA *et al.*, 2014).

2.4. Early computational studies of KIT mutants

The structural characterization of KIT mutants has been one of the focus of BiMoDyM group at ENS Cachan. In 2011, Elodie Laine has proven that KIT D816V mutation, positioned in the A-loop, induced a double effect – a local, manifested as the partial destruction of the small 3_{10} helix and a long-range structural reorganization of the JMR, followed by its release from the KD in the absence of extracellular ligand binding (LAINE *et al.*, 2011). After, it was evidenced that a communication route established between the distant A-loop and JMR in the native protein was disrupted in KIT D816V mutant (LAINE; AUCLAIR & TCHERTANOV, 2012).

Former member of our group, Isaure Chauvot de Beauchêne has also studied extensively the structural and dynamic effects of KIT mutants D816H/Y/N/V, V560G/D (CHAUVOT DE BEAUCHÊNE *et al.*, 2014) and later the S628N mutant (VITA *et al.*, 2014), by molecular dynamics simulations.

It was evidenced that as in D816V mutation studies, the A-loop mutations (D816H/Y/N) induce the inactive non-autoinhibited state of KIT evidenced by the the destabilization of the A-loop and the detachment of JMR from the TK domain. This effect conducts to deployment of the A-loop eventually leading to the constitutively active KIT state. The inactive non-autoinhibited state is not a suitable target for imatinib that inhibit the inactive autoinhibited state (Fig. 6). The JMR mutations (V560G/D) greatly impact the JMR binding to the kinase domain and facilitate its departure, favoring the non-autoinhibited state, whereas the inactive conformation of the A-loop is still conserved, which may facilitate inhibitors binding to the active site, thus increasing sensitivity to TKIs (Fig. 6) (CHAUVOT DE BEAUCHÊNE *et al.*, 2014).


A
WT
Inactive autoinhibited
X-ray (1T45) MD
JMR detachment
JMR binding
Inhibited
X-ray (1T46)
Inactive non autoinhibited
A-loop deployment
A-loop insertion
Active
X-ray (1PKG)

B

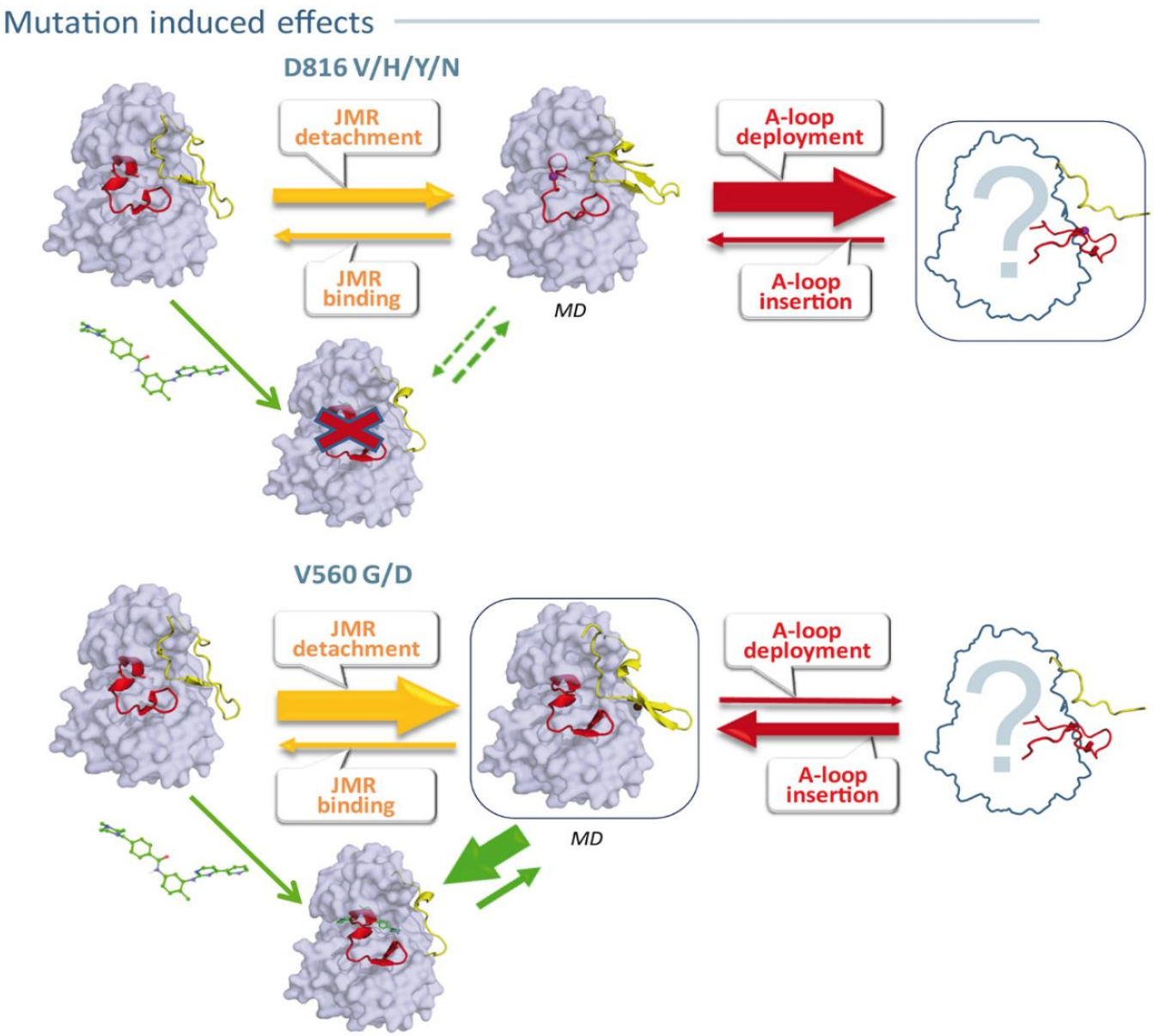

Mutation induced effects
D816 V/H/Y/N
JMR detachment
JMR binding
MD
A-loop deployment
A-loop insertion
V560 G/D
JMR detachment
JMR binding
MD
A-loop deployment
A-loop insertion

Figure 6: ***Proposed mechanisms of KIT activation by mutations****. The multi-states equilibrium of KIT cytoplasmic region in* KIT^{WT} *(A),* $KIT^{D816H/V/Y/N}$ *and* $KIT^{V560G/D}$ *(B: upper and lower panels). Each KIT conformation is represented as a molecular surface, except the JMR and the A-loop and imatinib drawn as cartoons and sticks respectively. In KIT mutants, the mutation position is shown by a ball. Equilibrium between two states is denoted by arrows of different thicknesses. (A) In the absence of SCF,* KIT^{WT} *is mainly in the inactive auto-inhibited state maintained by the JMR non-covalently bounded to the kinase domain. This state of KIT is the imatinib target. (B) Upper panel: The A-loop mutations (D816V/H/Y/N) induce the inactive non-auto-inhibited state of KIT evidenced by the JMR departure from the kinase domain. This effect conducts to deployment of the A-loop eventually leading to the constitutively active KIT state. The inactive non auto-inhibited state is not a suitable target for imatinib that inhibit the inactive auto-inhibited state. Lower panel: The JMR mutations (V560G/D) greatly impact the JMR binding to the kinase domain and facilitate its departure, favoring the non auto-inhibited state, whereas the inactive conformation of the A-loop is still conserved. The inactive non auto-inhibited state of KIT is more consented in KITV560G/D than in KITWT and especially in KITD816V/H/Y/N, leaded to the increased sensitivity of KITV560G/D to inhibitor compared to KITWT. In each panel, the most preferred state of KIT in the presence of imatinib is encircled. Figure and legend extracted from (CHAUVOT DE BEAUCHÊNE et al., 2014)*

The substitution S628N in KIT have a similar effect from the D816H/Y/N/V substitutions, accompanied of a higher flexibility of the Cα-helix (VITA *et al.*, 2014). D802V substitution in CSF-1R, although not frequently found in cancer, also triggers resistance to imatinib. The goals in this thesis were related to studying the structural and dynamical effects of this mutation and compare the results with KIT. After, we wanted to understand the inhibition mechanism of imatinib, complementing the mutation studies by investigating the receptor-inhibitor complexes by molecular dynamics.

2.5. Similarity between CSF-1R and KIT

CSF-1R and KIT have considerable sequence identity (54 %) and similarity (64%) and their auto-inhibited states display great structural similarities (RMSD is 1.14 Å) (WALTER *et al.*, 2007) (Fig. 7). Unlike the other type III RTK family members, the JM-S region of CSF-1R contains a unique conserved tyrosine (Y561, Y559 murine) (YU *et al.*, 2012). Consistent with the role of being a switch, Y559 is the first tyrosine to be phosphorylated and mutation of this residue to phenylalanine reduces significantly the *in vitro* kinase activity and markedly inhibits ligand-stimulated tyrosine phosphorylation in *vivo* (TAKESHITA *et al.*, 2007; YU *et al.*, 2008).

The sequence/structural similarity between the two receptors would permit us to expect the same effects for equivalent mutations, such as D816V (KIT) and D802V (CSF-1R), placed at the same point in structure (Fig. 7).



```
CSF-1R KYKQKPKYQVRWKIIESYEGNSYTFIDPTQLPYNEKWEFPRNNLQFGKTLGAGAFGKVVE 599
KIT    KYLQKPMYEVQWKVVEEINGNNYVYIDPTQLPYDHKWEFPRNRLSFGKTLGAGAFGKVVE 606

CSF-1R ATAFGLGKEDAVLKVAVKMLKSTAHADEKEALMSELKIMSHLGQHENIVNLLGACTHGGP 659
KIT    ATAYGLIKSDAAMTVAVKMLKPSAHLTEREALMSELKVLSYLGNHMNIVNLLGACTIGGP 666

CSF-1R VLVITEYCCYGDLLNFLRRKAEAMLGPSLSPGQDPEGGVDYKNIHLEKKYVRRDSGFSSQ 719
KIT    TLVITEYCCYGDLLNFLRRKRDSFI---CSKQEDHAEAALYKNLLHSKESSCSDS----- 718

CSF-1R GVDTYVEMRP-----VSTSSND-------SFSEQDLD----KEDGRPLELRDLLHFSSQV 763
KIT    -TNEYMDMKPGVSYVVPTKADKRRSVRIGSYIERDVTPAIMEDDELALDLEDLLSFSYQV 777

CSF-1R AQGMAFLASKNCIHRDVAARNVLLTNGHVAKICDFGLARDIMNDSNYIVKGNARLPVKWM 823
KIT    AKGMAFLASKNCIHRDLAARNILLTHGRITKICDFGLARDIKNDSNYVVKGNARLPVKWM 837

CSF-1R APESIFDCVYTVQSDVWSYGILLWEIFSLGLNPYPGILVNSKFYKLVKDGYQMAQPAFAP 883
KIT    APESIFNCVYTFESDVWSYGIFLWELFSLGSSPYPGMPVDSKFYKMIKEGFRMLSPEHAP 897

CSF-1R KNIYSIMQACWALEPTHRPTFQQICSFLQEQAQEDRRERDYTNLPSSS------------ 931
KIT    AEMYDIMKTCWDADPLKRPTFKQIVQLIEKQISESTNHI-YSNLANCSPNRQKPVVDHSV 956

CSF-1R --RSGGSGSSSSELEEESSSEHLTCCEQGDIAQPLLQPNNYQFC 973
KIT    RINSVGSTASSS--------------------QPLLVHDDV--- 977
```

Figure 7: ***Sequence and structural alignment between CSF-1R and KIT****. Above: Structure alignment of CSF-1R (PDB ID: 2OGV) and KIT (PDB ID: 1T45) are represented in cartoon, colored in orange and wheat, respectively. Below: Sequence alignment of CSF-1R and KIT, taking as reference the sequences deposited in the Uniprot database: P07333, residues 539-972 for CSF-1R; P10721, residues 546-976 for KIT. Identical residues are colored in grey; similar residues are colored in black. The boxes colored in blue, green and red represent, respectively the JMR, hinge and A-loop residues. The residue mutated in the 802/816 position is highlighted in pink.*

3. Allosteric regulation of RTKs

Allostery can be defined as the regulation of a specific protein`s activity by a local perturbation in structure that is propagated or affects the protein active site. Allosteric regulation has remained a central focus in biology due to the importance of understanding the fundaments of most processes beyond the molecular level, such as cellular signaling and disease (MOTLAGH *et al.*, 2014).

The first concepts of allostery are derived from the classic experiments of Changeux (CHANGEUX, 1961), which stated that two distinct sites within one protein, each binding different ligands, could interact despite being distant from each other in the molecular structure. In the absence of structural information, two dominant models for allostery were predominant: the `sequential` or KNF model (Koshland-Nemethy-Filmer) (KOSHLAND; NÉMETHY & FILMER, 1966) and the `symmetric` or MWC (Monod-Wyman-Changeux) model (MONOD; WYMAN & CHANGEUX, 1965).

MWC postulated the existence of two pre-existing quaternary states, tensed (T) and relaxed (R), whose equilibrium was shifted upon ligand-binding. The KNF model was based on the general notion of the inherent flexibility of the proteins, as `induced-fit` of a binding site in response to ligand. Both models are phenomenological and do not provide much insight into how the structure facilitates allosteric communication between sites (MOTLAGH *et al.*, 2014). With the development of structural biology, the model proposed by Perutz (PERUTZ, 1970; PERUTZ *et al.*, 1998) was the first to address allostery in terms of structural changes that could be gleaned through inspection of the high-resolution structure (Fig. 8). Advances in the structural biology techniques have permitted also to adapt the allosteric models to monomeric proteins, such as signaling proteins, in which the transmission of signals initiated at one functional surface to a distinct surface mediates downstream signaling.

Studies in other protein systems indicate that long-range interactions of amino-acids are also important in binding and catalytic specificity. Substrate recognition in the chymotrypsin family of serine proteases, the tuning of antibody specificity through B-cell maturation and the cooperativity of oxygen binding in hemoglobin all depend not only on residues directly

contacting substrate, but also on distant residues located in supporting loops and other secondary structural elements (SÜEL *et al.*, 2003).

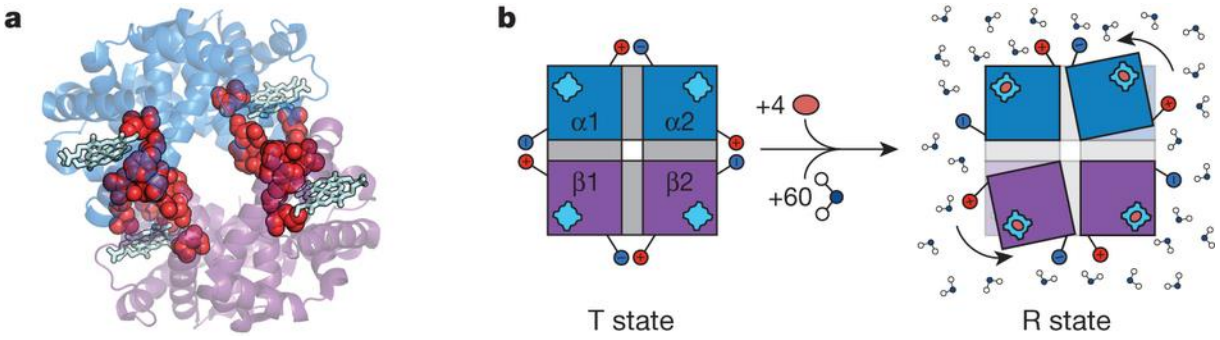


Figure 8: ***Schematic representation of the allosteric transition in hemoglobin****. (a) Ribbon diagram representation of tetrameric hemoglobin (PDB ID 1GZX). The proposed pathway responsible for the cooperative transition from tensed (T) to relaxed (R) is highlighted with red spheres and the heme groups are represented as light blue stick. (b) Allosteric transition of tetrameric hemoglobin, as proposed by Perutz* (PERUTZ, 1970; PERUTZ *et al.*, 1998)*. Tetrameric hemoglobin in the T state is depicted on the left with the two α-subunits (blue) and the two β-subunits (purple) each with their own heme group (light blue). Salt bridges, depicted as the red positive and blue negative charges, hold the molecule in the T conformation, and these salt bridges are released upon binding of oxygen (orange oval) in the transition to the R conformation (on the right) accompanied by a 15° turn of the subunits relative to each another. Also contributing to the equilibrium are 60 additional water molecules preferentially binding the R state. Extracted from* (MOTLAGH *et al.*, 2014)

Perturbation on a protein structure can occur not only by the binding of a substrate, inhibitor, or co-factor but also from a point mutation. These perturbations can be described in terms of signal propagation theory and molecular dynamics. Contrasting the model proposed by Perutz, there are evidences that have shown that allosteric coupling can be described by transmitted changes in protein dynamics as a consequence of a re-distribution of the protein conformational population (TSAI; DEL SOL & NUSSINOV, 2008), also reviewed at (MOTLAGH *et al.*, 2014). It suggests that allosteric information can result in global conformational changes or the modification of local atomic fluctuations. In either case, information transmission occurs through well-structured connectivity pathways or multiple dynamic micro-pathways in the protein residue network (DEL SOL *et al.*, 2009; KAR *et al.*, 2010).

A number of *in silico* techniques have been developed to predict the connectivity pathways that transmit the allosteric information among protein amino acids, mostly based on evolutionary conservation information (SÜEL *et al.*, 2003), native contacts within the

protein residue network (DIXIT & VERKHIVKER, 2011) or dynamical correlations from molecular dynamics simulations (MA & KARPLUS, 1998).

Recently, attempting to understand and characterize the allosteric communication in different forms of KIT (WT and D816V mutated), our group developed a modular network representation composed of *communication pathways* and *independent dynamic segments*, called MONETA (Modular NETwork Analysis) (LAINE; AUCLAIR & TCHERTANOV, 2012). MONETA consists on a mechanistic model of protein communication based on well-defined interactions by the introduction of concerted local atomic fluctuations.

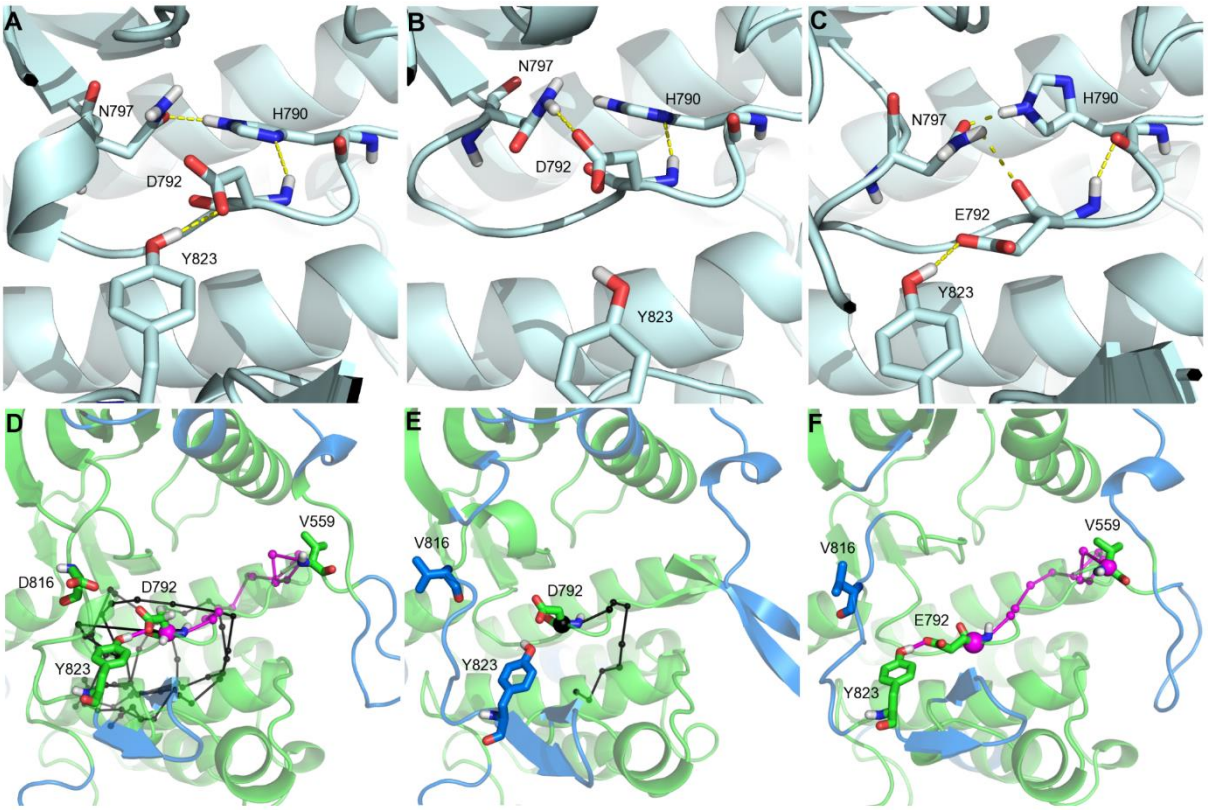


Figure 9: ***Interaction network and modular network representation in KIT cytoplasmic region****. Top: The interaction network between the A-loop tyrosine (Y823) and the catalytic loop residues (H790, D792 and N797) is depicted for WT KIT (A), the D816V mutant (MU) (B) and the D816V/D792E double mutant (C). The average conformation obtained from molecular dynamics is represented in pale cyan cartoons. H-bonds are displayed when their occupancy lies above 50% of the simulation time. Bottom: The modular network representations of the WT KIT (D), the D816V mutant (E) and the D816V/D792E double mutant (F) built by MONETA are depicted, focusing on the JMR, catalytic loop and A-loop regions. The average conformation obtained from molecular dynamics is represented in transparent cartoons. Communication pathways generated from residue in position 792 are displayed as chains of small black spheres connected together by black lines. The initial residue is highlighted by a bigger sphere centered on its Cα. The path linking the A-loop and the JMR through the catalytic loop is highlighted in magenta. Residues D/V816 and Y823 in the A-loop, D/E792 in the catalytic loop and V559 in the JMR are highlighted in licorice and labeled. Adapted from* (LAINE; AUCLAIR & TCHERTANOV, 2012)

It was evidenced by MONETA that there is a well-established communication between the A-loop and the distant JMR in the native protein and the D816V mutation provoked a disruption of such communication (Fig. 9). In addition, the communication was restored by *in silico* mutagenesis through a counter-balancing mutation (Fig. 9) (LAINE; AUCLAIR & TCHERTANOV, 2012). The communication patterns observed in native and mutated KIT correlated with their structural and dynamical properties observed by previous molecular dynamics simulations (LAINE *et al.*, 2011)

The description of networks established intra-protein represent an important step into understanding the allosteric regulation phenomena, proving that the perturbations in structure are not always obvious as a two-state model of protein extreme conformations, but much more subtle, through local perturbations in structure that can induce long-range effects as seen by the above example.

4. Molecular modeling of bio-macromolecules

Molecular modeling is a powerful approach for generation and analysis of three dimensional (3D) structures of biological macromolecules and it has been used to address a huge number of problems in structural biology in many ways. Modeling methods are often an integral component of structure determination by NMR spectroscopy and X-ray crystallography (FORSTER, 2002).

Crystallographic data can offer a great deal of information about the structure of bio-macromolecules, giving a first level of understanding about their functions. This information is, however, limited since each structure represents only one average conformation of the protein and this conformation depends on the crystallization conditions, as the name says, it gives a `frozen` static conformation. In addition, X-ray crystallography techniques have some drawbacks, such as being time-consuming and expensive. In addition, it is not a straightforward task to grow crystals from proteins. Moreover, to reach crystallization, the protein is generally modified by cleavage of the N- and C-terminal flexible parts, and/ or insertion/deletion of several residues or entire fragments. Such modifications produce ‘engineered’ crystal structures, which strongly differ from the original protein.

NMR techniques are able to offer additional information about protein function since they are able to describe intermediate conformational states. However, in structure determination by NMR methods, the main limitation is the size of the protein and its quantity, since large macromolecules require isotopic labeling to be studied (FORSTER, 2002).

Molecular modeling is one branch of the computational biology. It combines an ensemble of *in silico* methods designed to construct models of biomolecules in order to understand and predict their physico-chemical properties (structural, energetic and dynamical). These methods allow to describe biological processes at time scales that are not accessible by experimental methods, due to their limited resolution or their prohibitive cost in terms of time or biological material. Molecular modeling techniques address to structure prediction, protein folding/unfolding, molecular dynamics simulations, intra- and intermolecular interactions associated with structural and energetic characteristics, drug design, among others.

In this section, we will discuss the main modeling techniques used in this work: protein secondary and tridimensional structure prediction, molecular dynamics simulations, normal modes analysis, molecular docking and network analysis of intra-protein communication.

4.1. Protein structure and interactions

Proteins are polymers constructed from sequences of amino acids. They perform different kinds of functions and are essential to a proper functioning of living organisms. The basic blocks that form the proteins are the amino acids, and they are linked together via amide bonds, giving a polypeptide chain. The naturally occurring amino acids have a similar core but different side chains, which gives them a distinct nature in relation to size, polarity and hydrophobicity.

The regular intra-molecular interactions in a protein gives rise to some common structural motifs, such as α-helices and β strands. They constitute the *secondary structure* of a protein (the *primary* being the amino acid sequence and the *tertiary* the detailed three-dimensional conformation). Secondary structure elements can be connected by regions often referred as ‘loops’, since they adopt less regular structures.

Secondary and tertiary protein structures are held together through weak non-covalent interactions, such as hydrogen bonds, between polar residues, and van der Waals interactions. When present in a large number, these interactions contribute to stabilize the protein overall structure. The weak force of the intra-molecule interactions permit, also, the protein remodeling during a conformational change, such as the activation and deactivation of a cell-surface receptor protein, prompted by non-covalent ligand binding.

Another factor that contributes to protein stability is an interesting feature of water-soluble proteins: the *hydrophobic effect*, a consequence of the packing of hydrophobic amino acids, such as phenylalanine, tryptophan, valine and leucine, and the exposure of charged residues, as lysine, aspartate, glutamate and arginine in the protein surface (LEACH, 2001).

Not all proteins are water-soluble. For instance, membrane-bound proteins have a very different arrangement of amino acids in the membrane-spanning region, since the membrane environment is very hydrophobic and so, hydrophobic residues are often located on the outside, towards the membrane.

Regarding intermolecular interactions of protein with other entities – proteins, ligands, DNA/RNA, we observe the same rules, since we are dealing with atomic molecular systems. The next section will describe how the simulations programs treat the intra- (between bonded atoms) and inter- (non-covalent interactions) molecular interactions in order to describe as better as possible the psychical nature of these interactions.

4.2. Molecular mechanics, Quantum mechanics and Force fields

Due to the quantic nature of electrons movement in atoms, a consistent theory to describe the intermolecular interactions could only be derived from quantic-mechanics concepts. The Quantum Mechanics (QM) postulates that the dynamic properties of a quantum system can be described through predictions deduced from the wave propagation function, known as Schrödinger's equation. As an analog for the Newton's second law of motion, the time-dependent form of the equation can be described as:

$$i\hbar \frac{\partial}{\partial t}\psi = \hat{\mathrm{H}}\psi \qquad (1)$$

Where *i* is the imaginary unit, $\hbar$ is the Planck constant divided by 2π, ψ is the wave function that characterizes particle motion and $\hat{\mathrm{H}}$ is the Hamiltonian operator, which reflects the contributions of kinetic and potential energies to the total energy.

For complex reasons, it is not possible to solve the time-dependent Schrödinger's equation for a system composed of many hundreds of atoms. In this sense, many sorts of approximations have been developed to describe complex systems such as the ones composed by bio-macromolecules, which can reach thousands or millions of atoms. The Molecular Mechanics or Classical Mechanics is used for this purpose and many programs are based on a classical representation to calculate the physical properties of a system.

Classical mechanics describe the motion of bodies under the action of a system of forces based on the Newton`s second law of motion, $\vec{F} = m\vec{a}$, where $F$ is the force applied, $m$ the mass of the object and $\vec{a}$ the body`s acceleration. The potential energy of all systems in molecular mechanics is calculated using classical force fields (CFF). In computer simulations, the CFF is the description of a system formed by many particles by the superposition of simple terms, which describe the interaction between particles. An ensemble of empirical potential functions, adjusted by experimental data and theoretical (quantum) calculations is introduced into the potential function (VANGUNSTEREN & BERENDSEN, 1990).

The choice of the CFF must rely on the properties of the system to be analyzed. In common, they all possess a potential energy function that can be divided into two groups of terms: the first one represents the interactions between atoms bonded covalently (in general, harmonic representation of bonds, angles and improper angles deformation and bond torsions); the second group represents the interaction between non-bonded atoms (electrostatic, short range repulsion and attractive van der Waals forces).

The overall potential energy function is the sum of all terms described above. For example, the potential function for a molecular system composed of N atoms would be like:

$$E_{FF} = E_{bonds} + E_{angles} + E_{dihedral} + E_{improprer} + E_{vdW} + E_{elect} \qquad (2)$$

For each term in this equation, for a particular FF such as AMBER (HORNAK *et al.*, 2006) (version 9 and 10), one of the FF used in this work, we have:

$$E_{bonds} = \sum_{n=1}^{N_b} K_{b_{ij}}\left(r_{ij} - r_0\right)^2 \qquad (3)$$

whereN_b is the number of bonds in a molecule, $K_{b_{ij}}$ the bond force constant between two atoms (*i* and *j*) and r_0 the bond length in equilibrium (see Figure 10).

$$E_{angles} = \sum_{n=1}^{N_\theta} K_{\theta_{ij}}\left(\theta_{ij} - \theta_0\right)^2 \quad (4)$$

where N_θ is the number of angles in a molecule, $K_{\theta_{ij}}$ the angle force constant between atoms *i* and *j* and θ_0 the angle in the equilibrium (see Figure 10).

$$E_{dihedral} = \sum_{n=1}^{N_\varphi} \frac{V_{n,\varphi}}{2}[1 + \cos(n\varphi + \gamma)] \quad (5)$$

$$E_{improprer} = \sum_{n=1}^{N_{\varphi imp}} \frac{V_{n,\varphi imp}}{2}\left[1 + \cos\left(n\varphi_{imp} + \gamma\right)\right] \quad (6)$$

where $V_{n,\varphi}$ is the energy barrier of torsion, *n* being its periodicity and γ its phase (see Figure 10).

$$E_{vdW} = \sum_{i<j}^{N_{atoms}} f_{ij}^{L-J} \varepsilon_{ij}^* \left(\left(\frac{R_{ij}}{r_{ij}^{12}}\right)^{12} - 2\left(\frac{R_{ij}^*}{r_{ij}}\right)^6\right) \quad (7)$$

where R_{ij}^* and ε_{ij}^* are the Lennard-Jones parameters for an atom-pair *ij* (see Figure 10).

$$E_{elect} = \sum_{i<j}^{N_{atoms}} f_{ij}^{elect}\left(\frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_r r_{ij}}\right) \quad (8)$$

where ε_0 and ε_r are the permissiveness coefficient in vaccum and in the solution, respectively, and $q_i q_j$ the partial charges of atoms *i* and *j*.

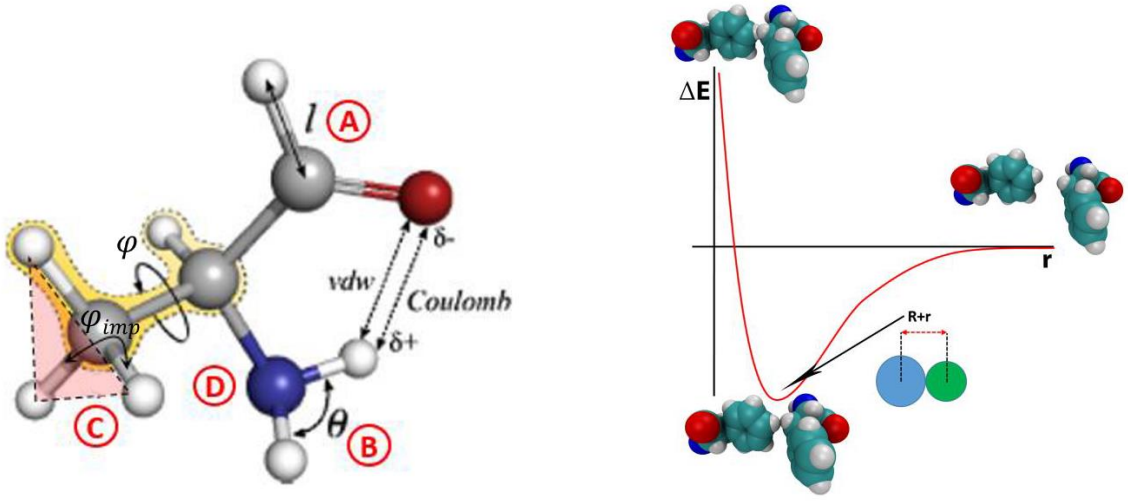


Figure 10: ***Schematic representation of the energy potentials related to bonded and non-bonded interactions.*** *Left. (A) Representation for the bond potential, where l represents the bond length. (B)*

Representation for the angle potential to any three bonded atoms, where ϑ is the angle between three consecutive bonds. (C) Improper dihedral potential, where φimp is the angle between two planes. This potential is important to keep, for ex, the planarity of benzene rings. (D) Dihedral (proper) potential, where φ is representing the angle with torsion freedom. Right. Lennard-Jones graphic representation with atomic representations of the sum of the atomic vdW radius. Figure adapted from (FERNANDES, 2014)

All the constants, partial atomic-charges, and other non-variable parameters have been characterized in details for each specific CFF and can be found in the literature or databases with specific values for proteins, nucleic acids, sugars, etc.

Depending on the force field, it can have other facultative terms relying either on the nature of analyzed system or of the simulation. For example, CHARMM CFF has an additional term to improve the conformational properties of protein backbones, called CMAP (MACKERELL; FEIG & BROOKS, 2004). It is important to mention that a CFF is defined not only by its functional form but also by its parametrization, in this case two CFF can have the same functional form but distinct parametrizations.

4.3. Minimization methods for geometry optimization

The optimization of molecular geometry is a technique applied to find the ensemble of the atomic coordinates in which the potential energy of molecule is at a minimum. Ideally, it would be the search for the global energy minimum, but due to the huge quantity of freedom degrees presented in bio-macromolecules, the exploitation of the whole multidimensional energy surface is practically impossible.

Therefore, the energy minimization methods search for a near local minima to avoid bond, angle and van der Waal tensions. A widely used method of energy minimization is the *steepest descent* method (BIXON & LIFSON, 1967). The *steepest descent* is a first-derivative method that converges slowly in the proximity of the energy minimum but is powerful to minimize conformations that are far from a local minima. Given the equation of the resultant force acting over each atom of the system:

$$\boldsymbol{F}_t = -\frac{\delta V(r_i)}{\delta r_i} \qquad (9)$$

derived from the total potential energy gradient, the steepest descent method can be defined as:

$$\boldsymbol{r}_{(i,n+1)} = \boldsymbol{r}_{i,n} + k_n \left(\frac{\boldsymbol{F}_{i,n}}{|\boldsymbol{F}_{i,n|}} \right) \quad (10)$$

where r gives the new position of atom *i* at step *n+1,* k_n is the step size adjustment parameter and $\frac{\boldsymbol{F}_{i,n}}{|\boldsymbol{F}_{i,n|}}$ is the single vector in the direction and sense of the resulting force over i in step n. The step or increment in the coordinates, $r_{(i,n+1)} - r_{i,n}$ of an atom i is given in the direction and sense of the resulting force over this atom.

In every step, the difference between the actual potential energy and the precedent one is verified; if the actual energy is lower than a stipulated value of energy, the calculation is stopped. Generally, setting this value to $\Delta V = 10^{-2} kcal/mol$, we can eliminate the deformity affecting bond lengths, angles and bad van der Waals and electrostatic interactions. The steepest descent method converges slowly in the proximity of local minima but is a good tool when you have a molecular configuration that is far from an energy minimum.

The *conjugate gradients* is also a first derivative method but it converges faster than the *steepest descent* method. The difference lies on the conception that in the *steepest descent*, both the gradients and the direction of successive steps are orthogonal; in *conjugate gradients*, the gradients at each point are orthogonal but the directions are *conjugate* (LEACH, 2001). The direction is computed from the gradient at the point and the previous direction vector. This property leads to a more direct path to the bottom of the energy potential, avoiding return over already traveled paths (Fig. 11).

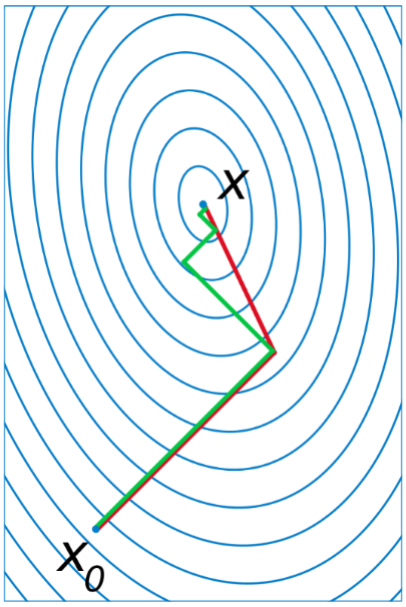


Figure 11: ***Graphical representation for the first derivative minimization methods****. The green and the red lines at the 2D representation of the energy well correspond, respectively, to the steepest descent and the conjugate gradients methods. Source: Wikipedia*

5. Protein structure prediction

5.1. Secondary structure prediction methods

Several bioinformatics methods are designed to predict the secondary structure elements of a protein starting from its primary sequence. Levinthal's paradox raised the question why and how the amino acid sequence can fold into its functional native structure among the infinite geometrically possible structures (LEVINTHAL, 1969). Proposed by Anfinsen's, first essays to answer this question have confirmed the hypothesis that the folding is a physical process that depends only on the specific amino acid sequence of the protein and the surrounding solvent (ANFINSEN, 1973).

The first-generation secondary structure prediction methods in the 1960s and 1970s were all based on amino acid propensities. The most popular second-generation methods in the early 1990s used propensities for segments of 3-51 adjacent residues, however the accuracy of prediction was restricted to ~60%, i.e. percentage of residues predicted correctly in one of the three states: helix, strand, and other (ROST, 2001).

The breakthrough of the secondary structure prediction methods came with the incorporation of information contained in multiple alignments and the use of evolutionary information, which in principle states that all naturally evolved proteins with more than 35%

pairwise identical residues over more than 100 aligned residues have similar structures (ROST, 1999). Evolutionary algorithms have increased the accuracy up to above 70%.

Another key factor was the introduction of position-specific profiles that described which residues could be exchanged against each other, containing crucial information about protein structure. The evolutionary divergence was the startup from many third-generation prediction methods. An example is the PSI-BLAST (ALTSCHUL, 1997), the gapped, profile-based and iterated search tool.

An alternative method widely used by many prediction tools is the application of Hidden Markov Models to represent sequence heterogeneity (ASAI; HAYAMIZU & HANDA, 1993). In a markovian sequence, the character appearing at position *t* only depends on the *k* preceding characters, being *k* the order of the Markov chain. Hence, a Markov chain is fully defined by the set of probabilities of each character given the past of the sequence in a k-long window: the transition matrix. In the Hidden Markov model, the transition matrix can change along the sequence. The choice of the transition matrix is governed by another markovian process, usually called the hidden process. In the case of secondary structure prediction, it is known that different classes have different sequence specificity, so, different Markov chains can be used to model different secondary structures.

A review written by Rost in 2001 (ROST, 2001) compares extensively the main methods used for secondary structure prediction at the time. The author concludes that growing databases and improved search techniques, predominantly through the iterated PSI-BLAST tool, yielded a substantial improvement in accuracy of the prediction. Another factor that increased confidence was the use of many different prediction methods.

Departing from that principle of combination of different methods, we selected a few tools available in web-based servers, to be used in this work.

In the next paragraphs, you will find a brief description of the tools used in this work.

a. GOR

GOR (GARNIER; GIBRAT & ROBSON, 1996) is an information theory-based method, based on probability parameters derived from empirical studies of known protein tertiary structures solved experimentally. It takes into account the individual amino acids propensities to form a

particular secondary structure, including a conditional probability of forming that structure given that its immediate neighbors have already formed it.

b. Jpred

Jpred (COLE; BARBER & BARTON, 2008) is a server that uses different algorithms to make the prediction. It uses the Jnet (CUFF & BARTON, 2000) algorithm to make the prediction of the secondary structure and solvent accessibility by combining BLAST (ALTSCHUL *et al.*, 1990), to search the protein sequence against sequences in the Protein Data Bank (PDB) (BERMAN *et al.*, 2000) and Uniref90 (SUZEK *et al.*, 2007); PSI-BLAST (ALTSCHUL, 1997), to make an alignment ; HMMer (EDDY, 1998), to construct an Hidden Markov model profile based on the alignment; and a Position-specific scoring matrix (PSSM) (JONES, 1999), output from PSI-BLAST.

c. SOPMA

SOPMA (GEOURJON & DELÉAGE, 1995) makes the prediction of the secondary structure based on the homolog method of Levin (LEVIN; ROBSON & GARNIER, 1986). Levin`s method divides the amino acid sequence in hepta-peptides and compares each one with a structural database of known protein structures, attributing a score to the comparison. SOPMA integrates to this method the exploitation of a multiple alignment, by applying the Levin`s method to a group of homologous sequences of known structures in order to optimize the prediction parameters which are specific to the target sequence.

d. SCRATCH

SCRATCH (CHENG *et al.*, 2005) combines machine learning methods, evolutionary information in the form of profiles, fragment libraries extracted from the PDB and energy functions to predict protein structural features and also tertiary structures.

e. NetSurfP

The NetSurfP (PETERSEN *et al.*, 2009) method consists of two neural network ensembles used to predict the secondary structure and the relative surface accessibility of an amino acid.

f. Psipred

Psipred (MCGUFFIN; BRYSON & JONES, 2000) uses a matrix PSSM calculated from a multiple alignment performed in a window of 15 residues by PSI-BLAST.

g. STRIDE

STRIDE (FRISHMAN & ARGOS, 1995) is a knowledge-based algorithm that assigns the secondary structure from atomic coordinates based on the combined use of hydrogen bond energy and statistically derived backbone torsional angle information.

5.2. Tridimensional protein structure prediction

In the past few years, there has been a lot of discussion about the vast wealth of data derived from the genome sequencing producing a “structural gap” since a minority of the protein sequences identified will have their structure solved by experimental techniques. Advances in DNA sequencing techniques are producing an unprecedented avalanche of new sequences (UniProt-Consortium, 2013), and it is obvious that it will be impossible to determine experimentally the structures of all proteins with the currently used techniques (SCHWEDE, 2013).

As discussed early in the begging of the section IV, the two main experimental techniques employed on the structure determination, have some drawbacks and limitations even nowadays. Fortunately, homologous proteins, which share sequence similarity, have similar or resembling three-dimensional (3D) structures. Based on this observation, methods for comparative modeling (or template-based modeling) of protein structures were developed in the two last decades, using the available experimental structure information to describe protein sequences non- structurally characterized. These techniques are nowadays matured into fully automated pipelines that, depending on some critical points, can provide reliable three-dimensional models accessible also to researchers which are non-specialists in structural or computational biology (Table 1, (SCHWEDE, 2013)). The critical aspects to be considered, such as sequence identity between template and target, quality of the experimentally solved structure, among others, can vary with the desired application of the final model.

Here we are going to discuss two-techniques of *in silico* 3D-structure prediction: **comparative modeling** and **template-free** protein modeling. The latter englobes the *ab-initio* and *de novo* prediction techniques.

Table 1: Commonly Used Tools and Services for Protein Structure Modeling and Prediction. Adapted from: (SCHWEDE, 2013)

Tool or Service	**Web Site**
Protein Model Portal	http://www.proteinmodelportal.org
Model Archive	http://modelarchive.org
HHpred	http://toolkit.tuebingen.mpg.de/hhpred
IMP	http://www.salilab.org/imp
IntFOLD	http://www.reading.ac.uk/bioinf/IntFOLD/
I-Tasser	http://zhanglab.ccmb.med.umich.edu/I-TASSER
ModBase	http://salilab.org/modbase/
Modeler/ModWeb	http://salilab.org/modeller/
Pcons.net	http://pcons.net/
PHYRE2	http://www.sbg.bio.ic.ac.uk/phyre2/
Robetta	http://robetta.bakerlab.org/
Rosetta	https://www.rosettacommons.org
SWISS-MODEL Repository	http://swissmodel.expasy.org/repository
SWISS-MODEL Workspace	http://swissmodel.expasy.org/workspace/

5.2.1. Comparative modeling

Comparative modeling of a protein consists in constructing an atomic-resolution model of a protein from its amino acid sequence (target) having as a template an experimentally obtained three-dimensional structure of another (homologues) protein. It relies on the identification of homologous proteins that resemble the structure of the query sequence.

The approach is based on the fact that the structure of a protein is more conserved than its primary sequence during the evolution, and small changes on the sequence, in general, lead to very subtle modifications on the structure (DA SILVA & BISCH, 2011; NAYEEM; SITKOFF & KRYSTEK, 2006).

The process of constructing a 3D model by comparative modeling is achieved in four following steps (Fig. 11): (i) template (s) identification; (ii) alignment between the template(s) and the target sequences; (iii) model construction; (iv) model validation (Fig. 12).

The template(s) identification of experimentally solved structures that can be used as a structural base for the target sequence modeling should be made with taking into account several aspects, such as structural knowledge, function similarity, sequence identity or evolutional correlation (DA SILVA & BISCH, 2011; HILLISCH; PINEDA & HILGENFELD, 2004; MARTÍ-RENOM *et al.*, 2000).

The model accuracy is generally related to the percentage sequence identity on which it is based. High accuracy comparative models are based on more than 50% sequence identity to their templates (BAKER & SALI, 2001), although nowadays is largely accepted that template-based prediction methods can be safely applied if target and template have more than 35% sequence identity for alignments of ~100 residues (LIU; TANG & CAPRIOTTI, 2011).

The accuracy of the model is a critical issue but model application has also to be considered. For example, if the aim is the drug design, target and template must have high similarity in the active site region (RMSD < 1.0 Å) (BAKER & SALI, 2001). However, the utility of low-accuracy models can be illustrated by fitting molecular models into electron microscopy maps, which allows the reconstruction of large biological machines (LASKER *et al.*, 2012; TOPF & SALI, 2005).

Once the template is chosen, the alignment between the sequences of target and template is done taking into account other factors, such as the secondary structure elements and the user's knowledge of the system. It is important to mention that a correct alignment is an essential step towards a good quality model. To improve the alignment coverage in case of low sequence identity (< 40%), multiples templates can be used in the alignment and model building (LIU; TANG & CAPRIOTTI, 2011). In this case, the manual inspection of the alignment is also recommended.

Different methods are used for the construction of the 3D model. The first and still used approach is by *assembly of rigid bodies* (BLUNDELL *et al.*, 1987). The approach is based on the dissection of the protein structure into conserved core regions, variable loops that connects them, and side chains coupled to the backbone.

A commonly used method applied in this thesis is through the *satisfaction of spatial restraints*. The idea is that structural features of conserved residues are similar. The evolutionary conservation is used as criteria to select and generate homology-based restraints

for the target structure using distances and angles between equivalent residues in the template. These restraints are normally supplemented with generic stereochemical restraints on bond lengths, bond angles, dihedral angles and non-bonded atom-atom contacts obtained from CFF. The model is then derived by minimizing the violations of all restraints (MARTÍ-RENOM *et al.*, 2000).

This methodology is implemented at the program *Modeller* (ESWAR *et al.*, 2008), and the approach used to derive the model is the real-space optimization method, based on the distance and dihedral angle restraints on the target sequence derived from the alignment with the 3D structure. The form of these restraints were obtained from a statistical analysis of the relationships between similar protein structures, based on a database of 105 family alignments (SALI & OVERINGTON, 1994). These relationships are expressed as conditional probability density functions and can be used directly as spatial restraints (ESWAR *et al.*, 2008).

Then, spatial restraints and CHARMM22 force field terms (MACKERELL *et al.*, 1998), which enforce the proper stereochemistry, are combined into an objective function, similar to those used in the molecular dynamics programs. The model is generated by optimizing the objective function in Cartesian space (MARTÍ-RENOM *et al.*, 2000), by using methods of conjugate gradients and molecular dynamics with simulated annealing (CLORE *et al.*, 1986). The restraints can be derived also from other experimental sources, such as NMR, cross-linking experiments, fluorescence spectroscopy, image reconstruction in electron microscopy, site-directed mutagenesis, and intuition among other sources (ESWAR *et al.*, 2008).

The last step, validation of the model, is a measure of the quality, which is highly correlated with the quality of resolution quality of the template structure (a criteria of good quality is a resolution equal or less than 2 Å) and the R-factor ($0.15 < R < 0.20$). The most commonly used way of estimating the quality is verifying the model's stereochemistry, using for example, the program Procheck (LASKOWSKI *et al.*, 1993). It evaluates the bond lengths, angles, rings planarity, chirality of the carbon atoms, side chain conformations, torsion angles from the main chain, and steric clashes between non-bonded atom pairs. The visualization can be done by the Ramachandran plot (RAMACHANDRAN; RAMAKRISHNAN & SASISEKHARAN, 1963) (Fig. 12).

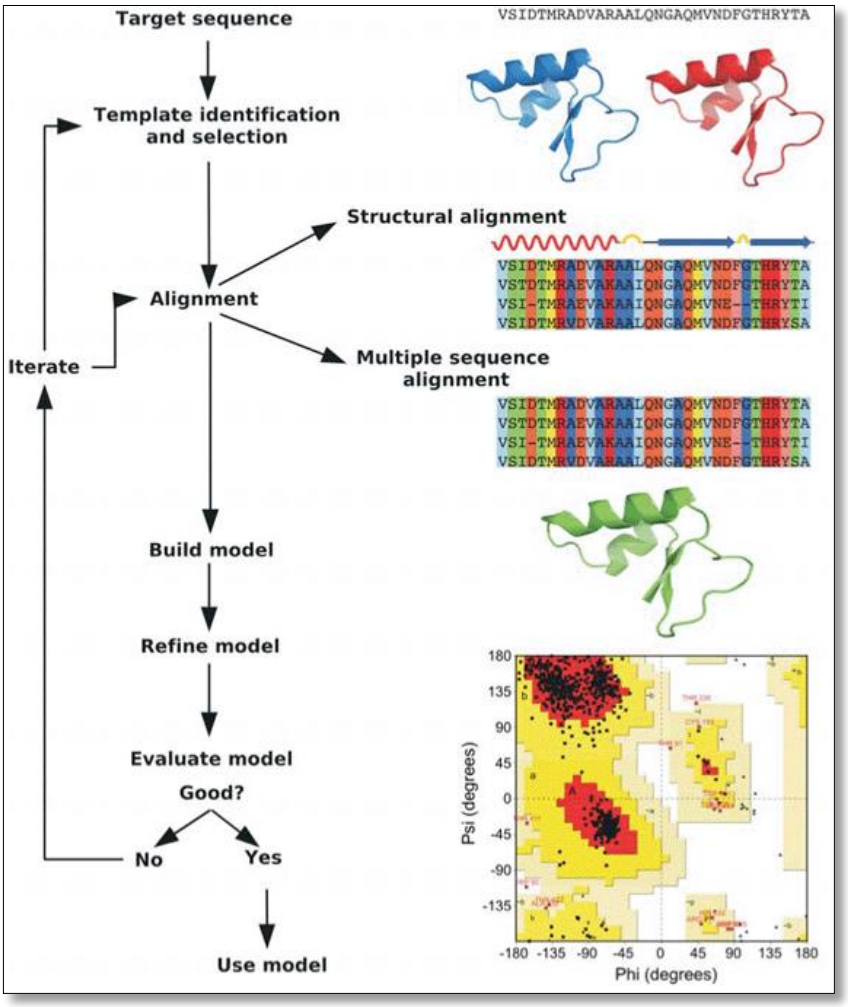


Figure 12: ***Comparative modeling process****. The first step involves the search and selection of homologue structures to be aligned with the target sequence. The alignment will serve as a backbone in which the model will be constructed. The next step involves the adjustment of the alignment using data derived from secondary structure prediction, for ex. After the model construction, its validation is done through using softwares as Watcheck* (HOOFT *et al.*, 1996) *and Procheck* (LASKOWSKI *et al.*, 1993)*, using the Ramachandran plot as criteria, for example. This graph permits to determine which torsion angles (ψ e ϕ) from the amino acid residues are correct, giving an idea about the precision of the model. Regions in red, brown and yellow represent, respectively, the favored, allowed and `generously allowed` defined by Procheck. Figure is reproduced from* (BISHOP; DE BEER & JOUBERT, 2008)*.*

The Ramachandran plot defines the residues that are placed in regions most `favorable` or `unfavorable` energetically. Glycine and proline residues can occupy unfavorable region due to their particular stereo chemical properties.

As an evaluation procedure, *Modeller* employs DOPE (SHEN & SALI, 2006), a method that use 3D profiles and statistical potentials to assess the compatibility between the sequence and the modeled structure. Other methods for evaluation based on the same

principle include VERIFY3D (LÜTHY; BOWIE & EISENBERG, 1992), PROSA2004 (WIEDERSTEIN & SIPPL, 2007), HARMONY (TOPHAM *et al.*, 1994), QMEAN local (BENKERT; BIASINI & SCHWEDE, 2011) and others.

5.2.2. Template-free protein modeling

Some approaches have been developed for *in silico* predicting 3D structure when there is no structural data available for homologue proteins or 'templates', or the identity degree between template and target is too low to be used in comparative modeling. *Ab initio* folding consists on the prediction of a protein`s structure and folding based only on its primary sequence, resulting in a novel fold (HARDIN; POGORELOV & LUTHEY-SCHULTEN, 2002).

The technique of threading, positioned between comparative modeling and *ab initio* methods, is an approach of fold recognition helping to construct a model of the target using a template structure of a protein that has little or no obvious sequence relation to the target protein.

These definitions, whether the technique is purely *ab initio* or uses some structural knowledge, are becoming more vague with the current programs, since the most successful methods utilize information from the sequence and structural databases in some form. Emerging from this discussion, the definition *template-free*, adopted by CASP (MOULT *et al.*, 2014), has appeared for this hybrid methods and they are highly performant in modeling at high resolution the structures of small proteins, composed of 25 - 100 amino acids.

The *de novo* protein structure prediction typically starts with predicting secondary structure and some other properties of the sequence. The programs can combine fragment-assembly based approaches (HANDL *et al.*, 2012) with folding simulation (PIANA; KLEPEIS & SHAW, 2014). Fragment assembly methods are based on the fact that the local folding of a protein is mainly resultant of local interactions, over the long-distance ones (FLOUDAS, 2007). This fact permits to restrict the search of the conformational space to model and evaluate it.

The programs select from a database an ensemble of fragments that covers the whole totality of the target sequence, and then optimize their assembly using molecular mechanics methods. ROSETTA (ROHL *et al.*, 2004) is one of the most performant programs for modeling small or big polypeptides, consisting of more than 25 amino acids.

Briefly, the ROSETTA structure prediction protocols generally begin with a low-resolution coarse-grained search of conformational space of the target, using a library of short peptide fragments (typically, of 3-9 residues long), constructed using the information from a secondary structure prediction method. The principle underlying a fragment selection is that the set of conformations sampled by a particular short sequence is likely to be reasonably well approximated by the set of conformations that similar sequence segments sample in known protein structures (GRONT *et al.*, 2011).

The conformational space spanned by these fragments is then searched using a Monte Carlo procedure with an energy function that favors hydrophobic burial and strand pairing and disfavors steric clashes. For each target sequence, a large number of decoy structures is generated using this protocol and then clustered; the five largest clusters are generally chosen as the predictions (LEAVER-FAY *et al.*, 2011). More information and the protocol details are reported in (LEAVER-FAY *et al.*, 2011; ROHL *et al.*, 2004).

The efficiency of ROSETTA method has been proved by the prediction of a new folding which was adopted by an artificial globular protein composed of 93 amino acids. The predicted folding was then validated experimentally by X-ray crystallography, with a structural deviation of 1.2 Å in respect to the predicted model (KUHLMAN *et al.*, 2003).

6. Molecular dynamics simulations

Molecular dynamics (MD) simulation is a determinist method that reproduces the 'real` dynamics of an atom-based system, from which time averages and properties can be calculated. Successive configurations of the system are generated by integrating the second Newton`s law of motion. The result is a trajectory that specifies how the positions and velocities of the particles in the system vary with time (LEACH, 2001).

The trajectory is obtained by solving the differential equations embodied in Newton`s second law ($\vec{F} = m\vec{a}$):

$$\frac{d^2x_i}{dt^2} = \frac{F_{x_i}}{m_i} \qquad (11)$$

This equation describes the motion of a particle of mass m_i along one coordinate represented by x_i, F_{x_i} being the force on the particle in the three directions and $\frac{d^2x_i}{dt^2}$ the acceleration.

The equations of motion are integrated using a *finite difference method*. This method is applied to generate MD trajectories with continuous potential models. The idea is that the integration is broken down into many small steps, separated in time by a fixed δt. The total force on each particle, in all-atom approximation, is calculated as the vector sum of its interactions with other atoms. From the force, the determined accelerations of the atoms are then combined with the positions and velocities at a time t to calculate the positions and velocities at a time $t + \delta t$. The force is assumed to be constant during the time step. The forces on the atoms in their new positions are then determined, leading to new positions and velocities at time $t + 2\delta t$, and so on.

The Verlet algorithm (VERLET, 1967) is probably the most widely used method for integrating the equations of motion in a MD simulation. It uses the positions and accelerations at time *t*, and the positions from the previous step, $\boldsymbol{r}(t - \delta t)$, to calculate the new positions at $t + \delta t$, $\boldsymbol{r}(t + \delta t)$:

$$\boldsymbol{r}(t + \delta t) = 2\boldsymbol{r}(t) - \boldsymbol{r}(t - \delta t) + \delta t^2 \boldsymbol{a}(t) \quad (12)$$

The velocities do not explicitly appear in the equation. It can be calculated in different ways, for example, a simple method is to divide the difference in position at times $t + \delta t$ and $t - \delta t$ by $2\delta t$:

$$\boldsymbol{v}(t) = [\boldsymbol{r}(t + \delta t) - \boldsymbol{r}(t - \delta t)]/2\delta t \quad (13)$$

Alternatively, the velocities can be estimated at the half-step, $t + \frac{1}{2}\delta t$:

$$\boldsymbol{v}\left(t + \tfrac{1}{2}\delta t\right) = [\boldsymbol{r}(t + \delta t) - \boldsymbol{r}(t)]/\delta t \quad (14)$$

The Verlet algorithm has several disadvantages: (i) a loss of precision, due to the addition of the small term $\delta t^2 \boldsymbol{a}(t)$ to the difference of two much larger terms, $2\boldsymbol{r}(t)$ and $\boldsymbol{r}(t - \delta t)$; (ii) difficulty in obtaining the velocities, since they are not explicit and cannot be available until the positions are computed at the next step; etc. To solve these issues, several alternative methods have been developed, such as the *leap-frog* algorithm (VAN GUNSTEREN & BERENDSEN, 1988a) and the *velocity Verlet* method (SWOPE, 1982).

The leap-frog algorithm uses the following relationships:

$$\boldsymbol{r}(t+\delta t) = \boldsymbol{r}(t) + \delta t \boldsymbol{v}(t+\tfrac{1}{2}\delta t) \quad (15)$$

$$\boldsymbol{v}\left(t+\tfrac{1}{2}\delta t\right) = \boldsymbol{v}\left(t-\tfrac{1}{2}\delta t\right) + \delta t \boldsymbol{a}(t) \quad (16)$$

The velocities $\boldsymbol{v}\left(t+\frac{1}{2}\delta t\right)$ are first calculated from the velocities at time $t-\frac{1}{2}\delta t$, and accelerations at time *t*. The positions are deduced from the velocities just calculated together with positions at time $\boldsymbol{r}(t)$. The velocities at time *t* can be calculated from:

$$\boldsymbol{v}(t) = \tfrac{1}{2}\left[\boldsymbol{v}\left(t+\tfrac{1}{2}\delta t\right) + \boldsymbol{v}\left(t-\tfrac{1}{2}\delta t\right)\right] \quad (17)$$

The velocities thus `leap-frog` over the positions to give their values at $t+\frac{1}{2}\delta t$. The positions then leap over the velocities to give their new values at $t+\delta t$, ready for the velocities at $t+\frac{3}{2}\delta t$, and so on. The main advantage of this approach over the Verlet algorithm is the explicit inclusion of the velocity which does not requires the calculation of the differences of large numbers. The main disadvantage is unsynchronized positions and velocities.

The velocity Verlet method gives positions, velocities and accelerations at the same time and does not compromise precision. The method is implemented as a three-stage procedure. In the first step, the positions at $t+\delta t$ are calculated according to the equation:

$$\boldsymbol{r}(t+\delta t) = \boldsymbol{r}(t) + \delta t \boldsymbol{v}(t) + \tfrac{1}{2}\delta t^2 \boldsymbol{a}(t) \quad (18)$$

using the velocities and the accelerations at time t. The velocities at time $t+\frac{1}{2}\delta t$ are then determined using:

$$\boldsymbol{v}\left(t+\tfrac{1}{2}\delta t\right) = \boldsymbol{v}(t) + \tfrac{1}{2}\delta t \boldsymbol{a}(t) \quad (19)$$

New forces are next computed from the current positions, thus giving $\boldsymbol{a}(t+\delta t)$. In the final step, the velocities at time $t+\delta t$ are determined using:

$$\boldsymbol{v}(t+\delta t) = \boldsymbol{v}\left(t+\tfrac{1}{2}\delta t\right) + \tfrac{1}{2}\delta t \boldsymbol{a}(t+\delta t) \quad (20)$$

The time required for integration is usually small compared to the other steps of calculation in a MD simulation. The most time-consuming part is the calculation of the force acting over

each atom. The integration method is concerned with the conservation of energy and momentum, being time-reversible, and should permit a long time step, δt, to be used.

The time-step is governed by the fastest degrees of motion in a system (such as bond vibrations). For example, bond vibrations involving hydrogens vibrate at the order of 10 fs. Due to these high frequencies, the time step in a molecular dynamics involving a biomolecule is generally set to 1-2 fs, and roughly up to 4s (HESS, 2008; SCHLICK, 2002), with the aid of algorithms to restrain the bonds geometry are applied, as SHAKE (KRAUTLER; VAN GUNSTEREN & HUNENBERGER, 2001) and LINCS (HESS *et al.*, 1997). Beyond a range of 5 fs, numerical instability sets in, and the coordinates and velocities of the trajectory grow significantly in magnitude (NYBERG & SCHLICK, 1992). If no restraint algorithm is applied, the time-step is set to 0.5 fs, in order to keep the trajectories stable.

The biological relevant motions, such as substrate catalysis, ligand recognition and folding occur on the microsecond to millisecond range, which is orders of magnitude higher than the possible time steps used in traditional MD simulations. Thus, simulating a medium-size protein requires months on computer time on a large distributed cluster to reach miliseconds of dynamics.

Hardware acceleration is one of the approaches used to reduce the computational cost of long MD simulations. Shaw's group has developed Anton, a specialized supercomputer where MD algorithms are implemented in hardware using application-specific instruction chips (ASICs) (SHAW *et al.*, 2007). Anton has proved improve speed up to 2 orders of magnitude over simulations of fully explicit solvated systems, in comparison with high performance computers. Another approach to increase speed that has become very popular due to its low cost is the use of graphical processing units (GPUs). GPUs development has been influenced by the entertainment industry of computer and video games. Their performance, combined with substantial increase of computing power can outperforms computing processing units (CPUs) (XU; WILLIAMSON & WALKER, 2010).

6.1. General MD protocol

The proteins dynamics are usually simulated in a solvated box, mimicking their natural environment. This model of simulation does not take into account the molecular crowding

related to real protein concentration inside the cell`s cytoplasm. The solvation can also be simulated implicitly, by modifying the dielectric constant of the medium. This type of solvation is less time-consuming in terms of computation but less precise in describing the dynamical features of the protein at atomic level.

The water models used in the MD simulations are developed for a specific force field and after they can be adapted to others force fields. For example, the TIP3P water model (JORGENSEN & JENSON, 1998), used in this thesis, is used to simulate the solvent explicitly in all-atom force fields, such as AMBER (CASE *et al.*, 2005) and CHARMM (MACKERELL; BANAVALI & FOLOPPE, 2000). In this model, the water molecule composed of three atoms (one atom of oxygen and two atoms of hydrogen) is kept rigid by a *pseudo-bond* between the two hydrogen atoms. Once the protein is placed inside the box, solvated by the water molecules, the neutrality of the system is established by adding counter-ions, to equilibrate the total charge of the system.

One of the goals in molecular simulations is to correlate the microscopic properties of the system with its macroscopic properties. The simulation of isolated systems, such as a protein in a solvated box, can suffer the named 'border effects', since the solvent molecules at the extremity of the box would interact with a smaller number of atoms than the molecules located at the center. To avoid or to minimize the border effects, the *periodic boundary conditions* (PBC) are used (CHEATHAM *et al.*, 1995). When using PBC, the box is replicated infinitely by translation on the three Cartesian directions (x,y,z) (Fig. 13). All the particles images are moving together but, in practice, only one of them is represented at the MD simulation: if an atom leaves the box at one extremity, one of its periodic images will enter by the opposite face.

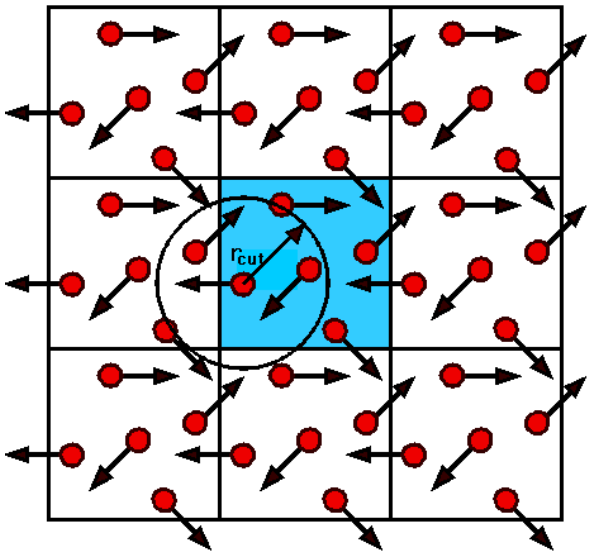


Figure 13: ***2D representation of the periodic boundary conditions in an infinite environment.*** *The central box is replicated on the three Cartesian coordinates and a cutoff radius, r_{cut}, can be used to restrict the region where the long-range interactions will be accounted for calculation during the MD simulation. Figure reproduced from* *http://wiki.cs.umt.edu/classes/cs477/index.php/Distance_Matrix#Periodic_boundary_conditions* *(accessed at 01/20/2015)*

The pseudo-infinite nature of the system implicates on the introducing of some approximations to treat the long-range interactions, number that would deeply increase. The introduction of a cutoff radius (Fig. 13) is used to minimize computational time, since the interactions outside the cutoff sphere are ignored. This value should be set to a value equal or less than half of the box dimension to assure that an atom will not interact with its own 'image'.

Currently, there are more refined methods to truncate the long-range interactions. For instance, *shift* or *switch* functions were developed to limit the drastic interruption in the treatment with the cutoff radius. The first one consists in adding corrective terms to the potential beyond the threshold value; the second modifies the potential between the threshold value and an intermediate distance (Fig. 14).

The contribution of van der Waals interactions at distances larger than 0.8 nm is close to zero, so the introduction of a cutoff radius does not compromise the accuracy. In contrast, the same principle will not be applicable to the electrostatic interactions, represented by the Coulomb potential in the force field. For instance, Coulomb interactions decay with 1/r; if the cutoff is too small, several interactions, that summed together have a significant contribution, would be eliminated.

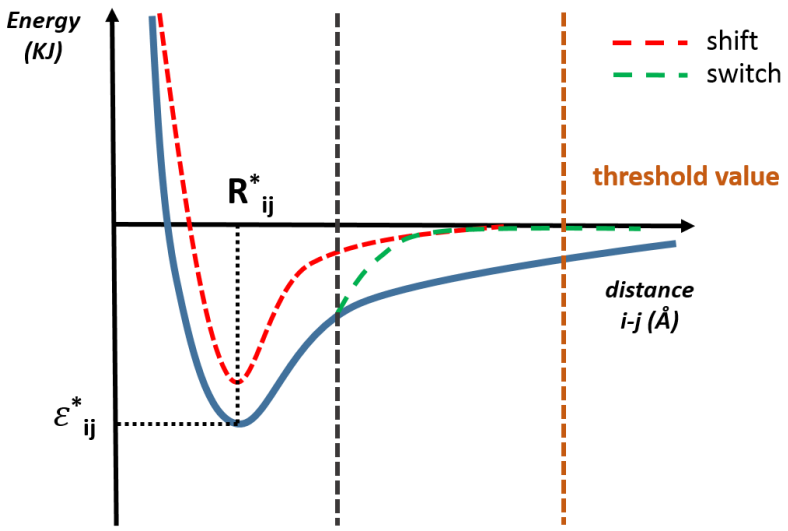


Figure 14: ***Cutoff methods for treating the long-range interactions****. Represented in the graph are the adjustments by shift (red) or switch (green) of the interaction energy, ε_{ij}^{*} (blue), between two atoms, i and j, in function of the interatomic separation, R_{ij}^{*}.*

In such cases, using the *Particle Mesh Ewald* (PME) (DARDEN; YORK & PEDERSEN, 1993) is preferable. It calculates electrostatic potential beyond a chosen cutoff distance by the solution of a Poisson-Boltzmann equation for a solute in a continuum solvent bath. The electrostatic potential within the cutoff distance is calculated as Coulomb forces. The method is based on the Ewald sum:

$$U_{ewald} = U_R + U_I + U_0 \qquad (21)$$

where U_R is the sum of interactions in short distance, U_I is the sum of interactions in long distance and U_0 is a corrective term. The first term is evaluated directly, while the second is approximated by a rapid Fourier transform in a grid where the charges are interpolated in each point (DARDEN; YORK & PEDERSEN, 1993).

Despite being widely used in the current MD simulations protocols, some issues have emerged related to the use of PME. For instance, it has been shown that the use of PME introduce artifacts that may bring unnatural bindings and overstabilization of the system (HÜNENBERGER & MCCAMMON, 1999). The overstabilization of the system, in the case of the proteins is not always appreciated since it could cause an artificial "freeze" of its geometry within a local energy minima.

Fadrna *et al*. (FADRNÁ; HLADECKOVÁ & KOCA, 2005) have shown that the cut-off treatment for electrostatic interactions have better reproduced the behavior of NMR solved peptides

structures (open conformation; C-terminus apart from the N-terminus), in comparison with PME, that led to an closed conformation of the peptide, with stable interactions between the charged peptide ends. Although controversial, this issue remains open and some authors point to the fact that incomplete sampling is more likely to affect the results to a larger extent than the artifacts induced by the use of Ewald sums (VILLARREAL & MONTICH, 2005). On the other hand, methods such as Reaction Field can cause an over flexibility of proteins and mask some effects, for example, small conformational changes due to mutations.

After the initial setup of the MD parameters described above, and *prior* the MD simulation itself, it is necessary to minimize the energy of the system. This is necessary to eliminate the physical constraints present in the crystal structure, which have highly elevated interatomic forces, and also to adapt the system to the selected force field used in the MD simulation. The process is done in several steps, commonly beginning by relaxing the water molecules, having the heavy atoms of the system with position restraints, and after, by relaxing the whole system - protein and solvent/counter-ion molecules.

As mentioned briefly in the previous section, some common algorithms are used to satisfy the bond geometry constraints in the MD simulations, such as SHAKE (KRAUTLER; VAN GUNSTEREN & HUNENBERGER, 2001) and LINCS (HESS *et al.*, 1997). The use of such methods remove the fastest degrees of freedom, allowing the increase of the time step in MD simulations and guaranteeing the energy conservation. Generally, this algorithms are applied at least to the bonds vibrations involving hydrogen atoms.

SHAKE is a two stage algorithm based on the Verlet integration scheme, and can be also used with its variant, the leap-frog algorithm. The Verlet leapfrog calculates the motion of the atoms assuming a complete absence of the rigid bond forces. The atom's positions at the end of this stage do not conserve the distance constraint required by the rigid bond and a correction is necessary; in the second stage, the length deviation obtained in the previous step is used to compute the constraint force needed to conserve the bond length (ALLEN & TILDESLEY, 1989). So, the SHAKE calculates the constraint force that conserves the bond lengths.

The LINCS algorithm is also suited for application with leap-frog or other Verlet-types integrators and can converge 3 to 4 times faster than SHAKE with the same accuracy (HESS *et*

al., 1997), besides being suited to parallelization in its most updated version (HESS, 2008). The method is built in the same linear approximations stated by SHAKE but is improved in some ways, for example, its iterations are applied to capture non-linear effects as bond rotations.

After the energy minimization, the system is submitted to an *equilibration MD simulation*, where it is gradually heated to achieve the goal temperature (physiologically relevant, usually of 310 K). After the equilibration, the system is ready for the *production MD simulation*, whose trajectory will be analyzed.

The equilibration and the production run are performed under specific thermodynamic ensembles in order to estimate the macroscopic properties of a system, such as temperature, pressure and volume, through the microscopic simulations. Three types of ensembles can be used in the MD simulations: the micro-canonical or *NVE*, where the number of particles, the volume and the energy are conserved; the canonical or *NVT*, where the number of particles, the volume and the temperature are conserved; and the isothermal-isobaric or *NPT,* where the number of particles, the pressure and the temperature are conserved (BROWN & CLARKE, 2006). The last two types, require a thermostat and the last one, a thermostat and a barostat, in order to control the temperature and the pressure of the system.

In the computer simulations, the temperature is computed from the kinetic energy of the system. The goal of a thermostat is not to keep the temperature constant but ensure that the average temperature of the system is correct. Also based on the kinetic energy of the system, the barostats are designed to modulate the pressure, usually by modifying the box vectors of the simulation cell and scaling the coordinates within the system. Such modulations can be applied uniformly (isotropic pressure coupling), independently in the x-y-z dimensions (semi isotropic pressure coupling) or independently in all directions (anisotropic pressure coupling) (VAN DER SPOEL *et al.*, 2005).

The Berendsen weak-coupling method is frequently applied in regulating the temperature and pressure of a system (BERENDSEN *et al.*, 1984). The method allows for exponential decay of an instantaneous value to the target value of pressure or temperature, so it does not generate a correct canonical ensemble. In equilibration phase, where the system is far from an equilibrium, the Berendsen weak-coupling can be applied to stabilize the system. In the data collection, a more accurate thermostat should be considered, such as the Nosé-Hoover

(CHENG & MERZ, 1996), that allows temperature to fluctuate about an average value, using a damping factor to control the temperature oscillation.

6.2. Analysis of MD trajectories

The production MD simulation contains information about the protein dynamics, which will be accessible after its careful analysis by analytical, statistical and/or graphical approaches. Some of the possible methods of analyses are discussed below.

6.2.1. Root Mean Square Deviation (RMSD) and Root Mean Square Fluctuation (RMSF)

Root Mean Square Deviation (RMSD) and Root Mean Square Fluctuation (RMSF) statistical techniques are used to study the system`s atomic coordinates deviation over the MD simulation. These parameters characterize the system stability. RMSD consists on the mean square deviation of the system`s particles in respect to a reference structure. Accordingly to the GROMACS manual, the calculation is done by the equation:

$$RMSD(t_1, t_2) = \left[\frac{1}{M}\sum_{i=1}^{N} m_i ||\boldsymbol{r}_i(t_1) - \boldsymbol{r}_i(t_2)||^2\right]^{\frac{1}{2}} \quad (22)$$

where, $RMSD(t_1, t_2)$ is the deviation of the atomic coordinates in time t_1 in relation to the coordinates in time t_2 (generally equal to 0 for the reference structure); $M = \sum_{i=1}^{N} m_i$, m_i is the mass of atom i and $\boldsymbol{r}_i(t)$ is the position of atom i at time t.

The RMSF computes the standard deviation of atomic positions in the trajectory in respect to an average conformation.

6.2.2. Convergence analysis

A convergence analysis can be performed on the trajectories using an ensemble-based statistical approach (LYMAN & ZUCKERMAN, 2006). The goal is to cluster all the conformations spanned by a protein in a MD simulation, according to their similarity between each other. The algorithm makes use of the global Cα atoms RMSD to discriminate representative MD conformations.

The procedure can be described as follows: (i) a set of *reference* structures are identified, (ii) the MD conformational ensemble is clustered into corresponding *reference groups*. Each *reference* structure is picked up randomly and associated with a bin of conformations distant by less than an arbitrary cutoff *r*. Once the reference groups are formed, they can be sub-divided in two ensembles: (a) conformations derived from the first half of the simulation time and (b) conformations derived from the second half of the simulation time. In a merged trajectory, that is, a combined trajectory derived from different MD simulations, (a) and (b) can refer to the first and the second trajectory, respectively.

A good convergence quality is assessed when each *reference* group is populated by conformations from the two halves of the trajectory (or MD simulations) at equivalent levels, meaning that every *reference* structure is equivalently represented in both parts of the trajectory (or in the MD simulations replica).

6.2.3. Secondary structure

Secondary structure analysis over a MD trajectory is generally performed using the DSSP method (KABSCH & SANDER, 1983), based on the identification of intra-backbone hydrogen bonds. Seven different patterns of secondary structure are distinguished: 3- and 5-*helices*, *α-helix*, *β- strand*, β-*bridge*, *turn* and *bend*. The *do_dssp* module, available at the GROMACS package (VAN DER SPOEL *et al.*, 2005), gives a temporal evolution of the secondary structure for each residue in the analyzed system.

6.2.4. Principal Components Analysis

Protein motion during a MD simulation is composed of movements at different frequencies and amplitudes. The movements of low frequency and, consequently high amplitude contribute the most to the global motion of a protein and represent the biologically relevant motions (AMADEI; LINSSEN & BERENDSEN, 1993). These low-frequency movements are generally collective and difficult to identify by simple visualization of the MD trajectory, and to discriminate among all the other motions such as the bonds and angle high-frequency vibrations of atoms.

Principal component analysis (PCA) is a widely used technique to retrieve dominant patterns and representative distributions from noisy data, such as a MD trajectory. The main

idea is to map the investigated protein system from a multidimensional space to a reduced space spanned by a few principal components (PCs) that can elucidate the main dominant features of a protein (GARCÍA, 1992; YANG *et al.*, 2009).

We use PCA to put in evidence the direction (eigenvectors) and amplitude (eigenvalues) along which the majority of the collective motions are defined. In practice, the calculations are performed on the backbone or Cα atoms positions recorded every 1 ps along the MD trajectories. Briefly, the method consists in calculating the covariance matrix *C* of all coordinates of a simulated protein. The covariance matrix indicates the common atom movements, translated in terms of eigenvectors e_r and eigenvalues λ_r:

$$C(i,j) = \sum_{r=1}^{3n} e_r(j).\lambda_r.e_r(i) \qquad (23)$$

Each eigenvector represents one *mode*, which describes a particular protein movement during the MD simulation. Each eigenvalue correlated to an eigenvector indicates the contribution of the mode to the protein dynamic behavior. Generally, the first modes account for most of the high amplitude protein movements.

In order to compare two eigenvectors issued from different simulations, it is necessary to make an approximation by overlap. The method consists in overlapping the subspace spanned by *m* orthogonal vectors *w1,…,wm* with a reference subspace spanned by *n* orthonormal vectors *v1, …,vn* and it can be quantified as follows:

$$overlap(v,w) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}(v_i.w_j)^2 \qquad (24)$$

The overlap will increase with increasing of *m* and will be 1 when set *v* is a subspace of set *w.*

6.2.5. Free energy of binding by the Molecular Mechanics (Poisson-Boltzmann /Generalized Born) Surface Area approach

The methods MM(PB/GB)SA, *Molecular Mechanics (Poisson-Boltzmann /Generalized Born)* Surface Area, permit the binding free energy calculation in a molecular complex by using an ensemble of protein conformations issued from a MD simulation of the protein complexed with a ligand (KOLLMAN *et al.*, 2000). The free energy of binding, ΔG_b is obtained accordingly to the equation:

$$\Delta G_b = \Delta E_{MM} + \Delta G_{sol} - T\Delta S \quad (25)$$

where ΔE_{MM} is the interaction energy derived from Molecular Mechanics, ΔG_{sol} the solvation free energy and $-T\Delta S$ represents the entropic contribution during binding. The term ΔE_{MM} is directly obtained from the MD data and corresponds to the sum of electrostatic and van der Waals energies between receptor and ligand:

$$\Delta E_{MM} = \Delta E_{elec} + \Delta E_{vdW} \quad (26)$$

The solvation free energy can be decomposed in two terms: electrostatic ($\Delta G_{sol/elec}$) and non-polar ($\Delta G_{sol/np}$):

$$\Delta G_{sol} = \Delta G_{sol/elec} + \Delta G_{sol/np} \quad (27)$$

The electrostatic component, $\Delta G_{sol/elec}$, can be obtained by either solving the linearized Poisson Boltzmann or Generalized Born equation (KOLLMAN *et al.*, 2000).

If PB is used, $\Delta G_{sol/elec}$ is obtained by the APBS program (BAKER *et al.*, 2001). APBS solves numerically the Poisson-Boltzmann equation and calculates the electrostatic energy. The Poisson-Boltzmann equation considers the solute (protein) charges explicitly and the solvent as a continuum medium, with an electrostatic potential that simulates a Boltzmann distributions for the ions surrounding the solute. The dielectric constant used for the solute can vary between 2 and 8, depending on the surface of the molecular system; for the solvent, the value is generally set to 80. The non-polar contribution, $\Delta G_{sol/np}$ is calculated as a function of the solvent accessible surface (SAS):

$$\Delta G_{sol/np} = \gamma(SAS) + b \quad (28)$$

where $\gamma = 0.00542\ kcal/mol\text{Å}^2$ and $b = 0.92\ kcal/mol$ (SANNER; OLSON & SPEHNER, 1996).

For reasons of computational efficiency, the program used to calculate the binding energy through the MM-PBSA approach, called *g_mmpbsa* (KUMARI *et al.*, 2014), uses the single-trajectory approach, in which is assumed that the conformational space accessible to the two binding species is unchanged on binding. This is accepted since MM-PBSA poorly predicts the binding energy associated with large conformational transitions (HOMEYER & GOHLKE, 2012). In addition, the program does not include the calculation of entropic terms

and therefore it gives the relative binding energy. Since we have used the same compound with proteins of similar structures and same binding mode, the entropy can be neglected. In the literature, some works point to the small net contribution of the entropy in terms of energy, leading to a non-significant improvement in the correlation with experimental values (BROWN & MUCHMORE, 2009; KUMARI *et al.*, 2014; RASTELLI *et al.*, 2010; YANG *et al.*, 2011).

7. Normal modes analysis

As discussed briefly at the section 4.2.4, the biomolecular structures can sample several degrees of freedom spanning from small bond vibrations to collective motions. The native structure of a protein is traditionally viewed as a single averaged structure, obtained by either experimental or theoretical techniques, while in reality the native structure is an ensemble of `micro-states` in dynamical equilibrium. These conformations share a common architecture and folding but they differ in their atomic coordinates, loop conformations, structure packing and even the position of structural sub-domains or domains (BAHAR *et al.*, 2010).

In the last years, there has been a significant increase in the number of studies that use the so called *elastic network models* (ENMs) and the *normal modes analysis* (NMA) to explore the protein structural dynamics (BAHAR *et al.*, 2010). NMA allows the description of low-frequency motions inaccessible to the time scale of most time-dependent methods, such as the short MD simulations (LEVITT; SANDER & STERN, 1985). In NMA calculations, is always assumed that the lowest frequency modes are the ones functionally relevant, because, they exist by evolutionary design rather than by chance (HAYWARD & DE GROOT, 2008).

NMA models the protein deformations as a harmonic oscillating system. The method considers a single structure as a system having a minimum in the potential energy surface of dimension *3N*, being *N* the number of atoms. In the vicinity of the energy minimum, one can consider the energy surface as quadratic, which allows the description by a Hessian matrix, ***F***, whose elements are derived from the second derivate of the energy function in respect to the atomic coordinates. The diagonalization of the Hessian matrix provides the vectors and frequencies of the normal modes (BROOKS & KARPLUS, 1983).

At a given temperature, the Hessian, ***F***, is inversely proportional to the covariance matrix of the atomic displacements, σ:

$$\boldsymbol{F} = k_B T \sigma^{-1} \quad (29)$$

where k_B is the Boltzmann constant, T the absolute temperature and each element of σ is defined as in:

$$\sigma_{ij}^{NM} = k_B T \sum_{l=1}^{3N-6} \frac{\alpha_{il}\alpha_{jl}}{\omega_l^2} \quad (30)$$

where α_{il} is the ith component of lth vector, ω_l is the frequency of normal mode lth and the sum is made over the internal normal modes in the order of $3N-6$ degrees of freedom (BATISTA *et al.*, 2010; KARPLUS & KUSHICK, 1981).

Comparing NMA and PCA, they share some similarities and differences. NMA is based on single-structures and does not require an MD simulation, depending only on the topology of protein native contacts. Similar to PCA, NMA rests on the assumption that major collective modes of fluctuation dominate the functional dynamics. In contrast, PCA does not rest on the assumption of a harmonical potential, since eigenvectors are not obtained by the diagonalization of the Hessian matrix, which contains derivatives of the forces with respect to every coordinate as elements. So, the large concerted motions from a MD trajectory have no restrictions to the shape of the potential energy function (VANAALTEN *et al.*, 1997).

In a relatively recent study, authors have compared both methodologies using as object of study the GroEL chaperone protein (SKJAERVEN; MARTINEZ & REUTER, 2011). The results showed a qualitatively good agreement between the movements described by the first five modes obtained with the three different approaches used: PCA, all-atoms NMA and coarse-grained NMA, where a simple representation of the protein chain is used. These results point to an advantage of the NMA approaches, since they have a reduced computational cost in comparison with an all-atom MD simulation, depending on the used platform. In addition, it was evidenced that the PCA presented a poor reproducibility comparing individual MD runs, which is a main disadvantage for the method.

8. Modular network analysis with MONETA

Discussed briefly in the section 3 of Introduction, MONETA is a program developed with the aim of building a modular network of the protein that describes the allosteric coupling in a rational way (ALLAIN *et al.*, 2014; LAINE; AUCLAIR & TCHERTANOV, 2012).

The modular network representation of the protein is composed of clusters of residues representing *Independent Dynamic Segments* (*IDSs*) and chains of residues representing *Communication Pathways* (*CPs*) (Fig. 15) (ALLAIN *et al.*, 2014; LAINE; AUCLAIR & TCHERTANOV, 2012). This representation is derived from the protein topology and the inter-residue dynamical correlations calculated on a conformational ensemble obtained by MD simulations. *CPs* are generated based on the *communication propensities* (CHENNUBHOTLA; YANG & BAHAR, 2008) between all protein residues.

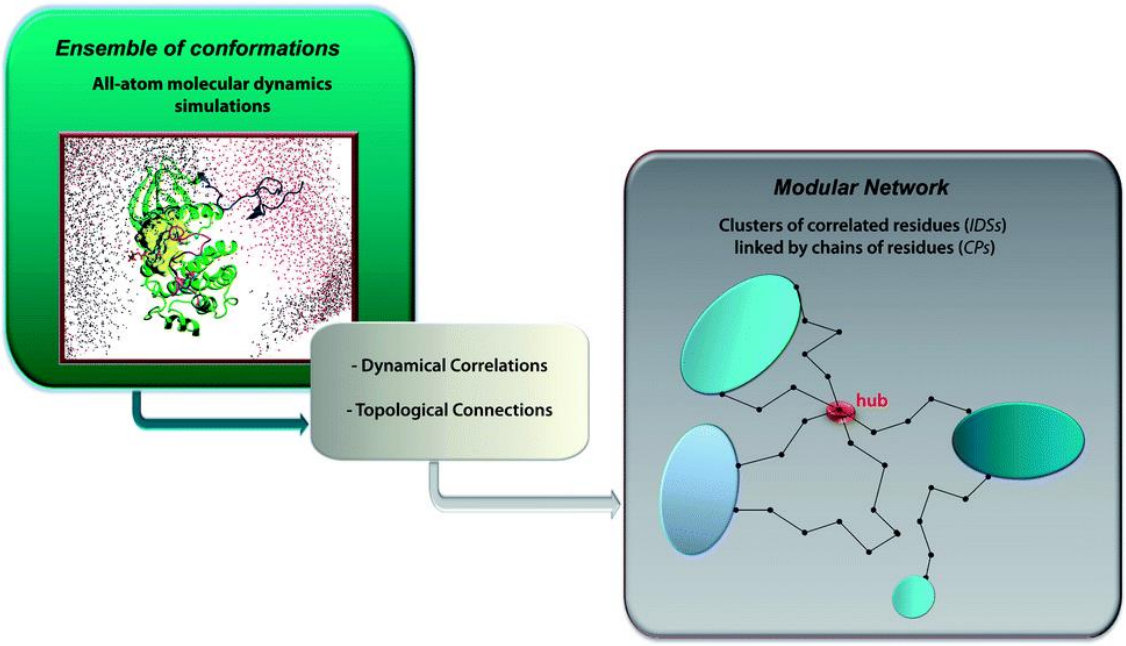


Figure 15: ***Schematic representation of the Modular MONETA's general principle****. A modular network representation composed of clusters of residues and chains of residues is built from the dynamical correlations and topology calculated from a protein conformational ensemble. In MONETA, residue clusters or modules are delineated as independent dynamic segments (IDSs) as they represent the most striking features of the protein local dynamics. Chains of individual residues are designated as communication pathways (CPs) as they represent well-defined connectivity pathways along which interactions can be mediated at long distances in the protein. Information is propagated through IDSs via the modification of the local atomic fluctuations and through CPs via well-defined interactions. The highly connected residues, at the junction of many pathways, can be considered as "hubs" in the protein network. Figure extracted from* (ALLAIN *et al.*, 2014).

MONETA analyses and extracts data from MD trajectories to infer the topological connections (residues interactions) and the dynamical correlations between residues or domains (Fig. 15). The analysis is consisted of three steps (ALLAIN *et al.*, 2014):

i. Identification of the IDSs, the protein regions displaying the most striking features of the protein's local dynamics;
ii. Detection of *CPs* linking the IDSs through non-bonded interactions between residues;
iii. Visualization of IDSs and CPs in a communication profile of the protein.

IDS constitute clusters of residues in which the atomic fluctuations are highly concerted within each cluster, although independent from the rest of the protein. IDSs are identified by a statistical technique called Local Feature Analysis (LFA) (PENEV & ATICK, 1996), adapted for analysis of the atomic coordinate fluctuations from MD simulations (ZHANG & WRIGGERS, 2006).

The ability of the protein residues to communicate efficiently is evaluated by using the measure of communication propensity (CHENNUBHOTLA; YANG & BAHAR, 2008). The communication between two residues is estimated by their commute time, expressed as the variance of their inter-residue distance over MD trajectories (DIXIT & VERKHIVKER, 2011). Chains of residues interacting by pair and displaying high communication propensities between them would represent pathways of well-defined interactions through which information would be transmitted efficiently. Such chains of residues are denoted as *CPs*. More details in the theory can be found at (ALLAIN *et al.*, 2014; LAINE; AUCLAIR & TCHERTANOV, 2012).

MONETA package performs automatically all the computational steps and analysis through a python scripting interface to the softwares R (IHAKA & GENTLEMAN, 1996), PyMOL (DELANO, 2004) and Gephi (MATHIEU BASTIAN, 2009). Somme additional softwares are required to analyze the MD trajectories: *ptraj* module of AmberTools (CASE *et al.*, 2005), *g_mdmat* module of GROMACS (VAN DER SPOEL *et al.*, 2005), HBPLUS and HBADD (MCDONALD & THORNTON, 1994). The workflow can be seen at Figure 16. MONETA outputs contain a 2D network graph, which translates the connectivity groups and pathways and a 3D representation of the CPs at the atomic level using PyMOL (Fig. 16).

Since it is a very recent developed technique, examples of applications in the literature are found only within our group. A very interesting result was obtained with MONETA by studying the allosteric propagation effects induced by the D816V mutation in KIT (LAINE; AUCLAIR & TCHERTANOV, 2012), this particular case being already used as example in section 3 of the Introduction chapter.

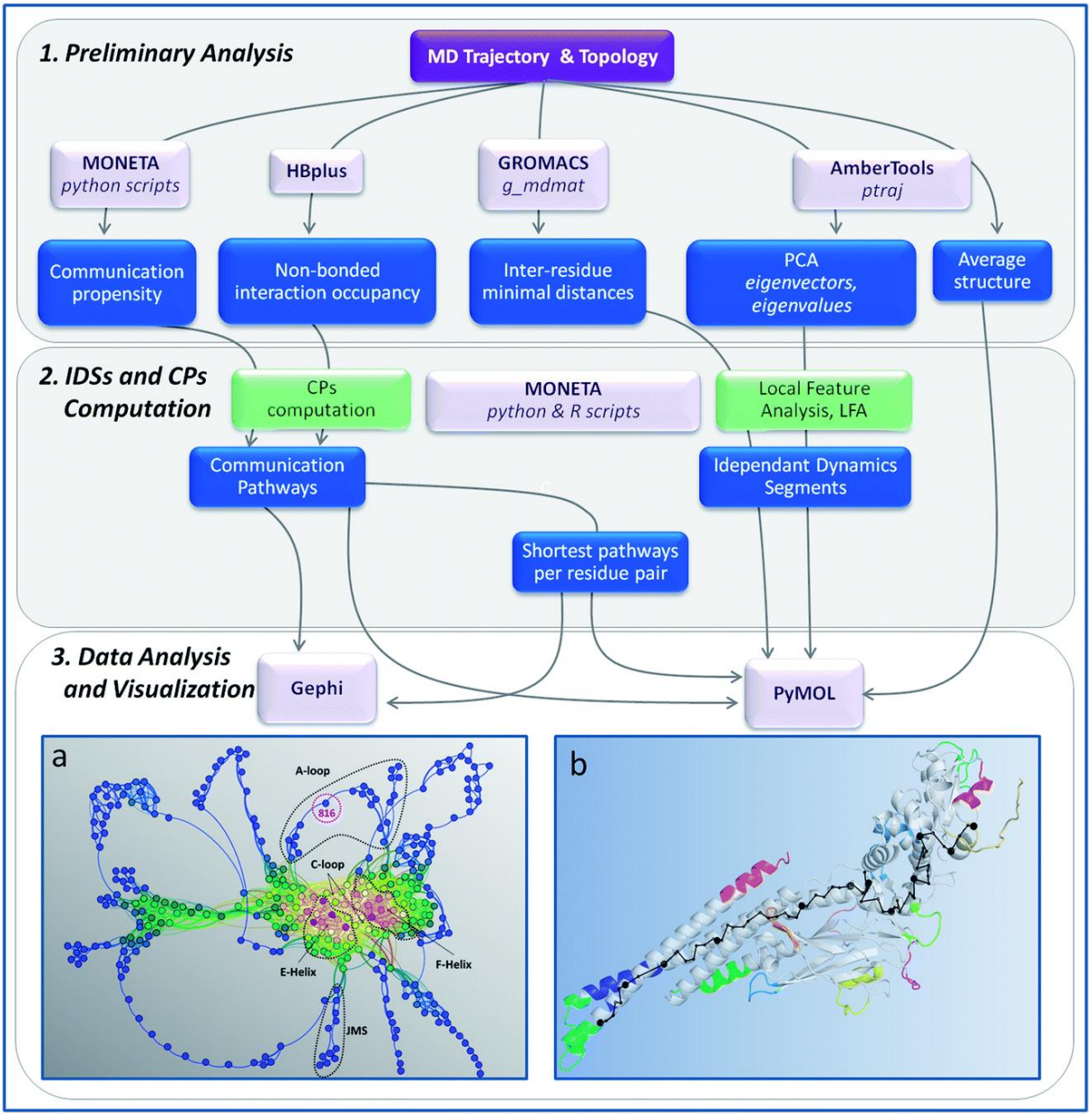


Figure 16: ***Overview of the major analysis steps in the MONETA workflow****. Each step of the MONETA procedure is delimited by an icon. The required inputs, parameters, outputs and scripts are identified by colors: initial mandatory inputs in purple, outputs in blue, MONETA computation steps in green, software and program in grey. Step 3 is illustrated by a 2D graph of the communication landscape in KIT (a) and by 3D representation of communication pathway in STAT5 (b). 2D and 3D graphs drawn with GEPHI and PyMOL modules incorporated in MONETA. Figure extracted from* (ALLAIN *et al.*, 2014)

9. Molecular docking

Molecular interactions between two or more biological entities (protein-protein, enzyme-substrate, protein-nucleic acid, drug-protein, etc) play important role in nearly all crucial biological processes, such as signal transduction, transport, cell regulation, gene expression control, enzymatic reactions and its inhibition, antibody-antigen recognition, and even the assembly of multi-domain proteins (LI; HAN & YU, 2013).

These interactions lead to the formation of stable or meta-stable receptor-ligand interactions that are essential to perform their biological functions. Most of the time, it is difficult to obtain the structure of all molecular complexes by experimental techniques. Consequently, some computation techniques (e.g. molecular docking) have been developed and optimized to study these biological phenomena.

Molecular docking is a widely used computer simulation procedure designed to predict the conformation of a receptor-ligand complex, where the general term 'receptor' refers usually to a protein or nucleic acid and 'ligand' is either to a small molecule or to another protein. One of the key applications of docking simulations is the *structure-based drug design*, e.g. the rational development of new compounds suitable to be used as medicaments, since the knowledge of the binding mode of a ligand can give insights of how it could be improved to better bind to the protein active or binding site. Another application is the *virtual screening*, where a database of thousands or millions of known compounds can be tested against a protein or enzyme of interest, in order to discard the compounds that would not possibly interact, or interact weakly with the target, saving time and unnecessary experiments.

The first docking programs were based in the model of molecular recognition proposed by Emil Fischer in 1894, named "Lock-and-key", where the protein binding site should be exactly complementary to the ligand shape, as a key in a locker (FISCHER, 1894; LICHTENTHALER, 1995). In this approximation, the programs used to treat the protein as rigid bodies. Afterwards, the programs were optimized to introduce flexibility degrees for the receptor, based on the concept of induced fit, proposed by Koshland in 1958 (KOSHLAND, 1958). In this model, both receptor and ligand structures adapt to each other during the binding.

9.1. Receptor flexibility

Nowadays, most docking programs open the possibility of treating the receptor as flexible, although it is still challenging due to the large size of the proteins and the many degrees of freedom of its components (atoms, residues and structural elements or domains), increasing the complexity of the problem. To address this, four different methodologies, reviewed at (HUANG & ZOU, 2010) are generally employed :

(i) *soft docking*, in which the receptor flexibility is considered implicitly by softening the van der Waals interactions during the calculations;

(ii) *side-chain flexibility*, in which the protein backbone is kept fixed and side-chain conformations are sampled;

(iii) molecular relaxation, a method that consists in relaxing or minimizing receptor-ligand complexes initially formed by rigid-body docking, taking in consideration the backbone besides the side-chain movement;

(iv) docking of multiple structures consists of utilizing an ensemble of protein structures to represent different possible conformational changes of the receptor.

The docking simulations presented in this work were performed using the *Induced-Fit* (IF) approach (SHERMAN *et al.*, 2006), available at Maestro (Schrödinger suite). IF protocol uses Glide docking program (FRIESNER *et al.*, 2004; HALGREN *et al.*, 2004) and Prime structure prediction program (JACOBSON *et al.*, 2004) to exhaustively consider the possible binding modes of a ligand and the associated conformational changes within receptor active or binding sites.

IF begins by docking the active ligand with Glide, the actual docking program for ligand-receptor docking in Maestro. In order to generate a diverse ensemble of ligand poses, the procedure uses reduced van der Waals radii and an increased Coulomb-vdW cutoff, and can temporarily remove highly flexible side chains during the docking step. For each pose, a Prime structure prediction is then used to accommodate the ligand by reorienting nearby side chains. These residues and the ligand are then minimized. Finally, each ligand is re-docked into its corresponding low energy protein structures and the resulting complexes are ranked

according to *GlideScore,* the empirical scoring function used to approximate the ligand binding free energy ("Small-Molecule Drug Discovery Suite: Induced Fit Docking protocol", 2014).

Before the application of the IF protocol, some preparation steps are required for the protein and the ligand. In Maestro, this tasks are performed by the *Protein Preparation Wizard* and the *LigPrep* programs. Schrodinger's *Protein Preparation Wizard* can solve common issues related to x-ray crystallographic structures by adding missing hydrogen atoms, incomplete side chains and loops, ambiguous protonation states and flipped residues in an automated fashion. *LigPrep* is devoted to prepare the ligand structure by optimizing the stereo chemical and ionization variations of the molecule and energy minimization. Together, this two programs assure a correct starting configuration for the docking procedures.

9.2. Ligand sampling

Ligand flexibility can be sampled accordingly to three types of search algorithms (BROOIJMANS & KUNTZ, 2003; HUANG & ZOU, 2010): *systematic, stochastic* and deterministic (e.g., energy minimization or molecular dynamics). The first tries to explore systematically every degree of freedom in a molecule. The number of combinations can be huge with the increase of rotatable bonds. Therefore, to make the process more practical, the geometrical/chemical constraints are applied to the initial screening of ligand poses, and the filtered ligand conformations are further subjected to more accurate refinement/optimization procedures. Glide (FRIESNER *et al.*, 2004; HALGREN *et al.*, 2004) is a typical example of this sampling method. Many algorithms use an incremented construction of the ligand, such as the one employed in FLEX (RAREY *et al.*, 1996), that is, dividing the ligand into a rigid core and the flexible side chains, defined by perception of the bonds possible to rotate. The rigid core is anchored first, and the flexible parts are added sequentially, as a systematic screening of the torsion angles. Another systematic approach would be the employment of pre-generated conformations library, such as used in FLOG (MILLER *et al.*, 1994), developed by the Merck research laboratories.

Stochastic methods consist in making random alterations on the structure of a ligand or a sample of ligands. The resulting structure is evaluated accordingly to a probability function. The most commons methods are Monte Carlo and evolutionary algorithms, represented

mostly by the genetic algorithms (HALPERIN *et al.*, 2002; KITCHEN *et al.*, 2004). For example, programs such as LigandFit (VENKATACHALAM *et al.*, 2003) and MoDock (GU *et al.*, 2015) use Monte Carlo; GOLD (VERDONK *et al.*, 2003) and AutoDock (MORRIS *et al.*, 1996) have genetic algorithms among their setup options.

The most popular deterministic method is the simulation by molecular dynamics. However, molecular dynamics simulations under physiological temperatures are incapable of crossing high energy barriers in accessible time scales, which would accommodate the ligand in global energy minimums (BROOIJMANS & KUNTZ, 2003). One alternative solution is coupling the molecular dynamics with the simulated annealing method, where the different freedom degrees of the system are coupled to different temperatures (DINOLA; ROCCATANO & BERENDSEN, 1994).

9.3. Scoring functions

The scoring function is a key element of each docking program, because it affects directly the accuracy of the prediction. Numerous scoring functions have been developed in the last decades and can be re-grouped into three principle categories: *force field*, *empirical* and *knowledge-based*.

Force field scoring functions are based on decomposition of the ligand binding energy into individual interaction terms such as van der Waals (VDW) energies, electrostatic (Coulomb) energies, bond stretching/bending/torsional energies, using a set of parameters derived from force-fields such as AMBER (CASE *et al.*, 2005) or CHARMM (BROOKS *et al.*, 2009). One major issue is to take into account the solvation and entropic terms. To reduce the computational cost caused by treating water molecules explicitly, the simplest method is the use of a distance-depending dielectric constant $\varepsilon(r_{ij})$ to consider implicitly the solvent effect, such as the CFF function below described for the program DOCK (HUANG; GRINTER & ZOU, 2010):

$$E = \sum_i \sum_i \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} \right) \quad (31)$$

where r_{ij} is the distance between protein atom *i* and ligand atom *j*, A_{ij} and B_{ij} are the VDW parameters, and q_i and q_j are the atomic charges, $\varepsilon(r_{ij})$ is usually set to $4r_{ij}$, reflecting the

screening effect of the water molecules on the electrostatic interactions. The Poisson-Boltzmann/surface area (PB/SA) models and the generalized-Born/surface area (GB/SA) are other examples of implicit solvent models (HUANG & ZOU, 2010).

In empirical scoring functions, the docking binding energy is calculated as a sum of weighted empirical factors, such as VDW, electrostatic, hydrogen bonding, desolvation term, entropy component, etc.:

$$\Delta G = \sum_i W_i \,.\, \Delta G_i \qquad (32)$$

where ΔG_i represent individual empirical energy terms, and the corresponding coefficients W_i are determined by reproducing the binding affinity data of a training set of protein-ligand complexes with known three-dimensional structures, determined experimentally.

Compared to the force field scoring functions, the empirical scoring functions are normally much more computationally efficient. The increased number of experimentally determined crystal structures of diverse protein-ligand complexes has made it possible to develop empirical functions by training on the binding constants of thousands of protein-ligand complexes. GlideScore (FRIESNER *et al.*, 2004; HALGREN *et al.*, 2004) is an example of this kind of functions.

Knowledge-based functions are derived directly from the structural information in experimentally determined protein-ligand complexes. The principle behind knowledge-based functions is the potential of mean force, which is defined by inverting the Boltzmann relation:

$$\mathfrak{w}(r) = -k_B T ln[\rho(r)/\rho^*(r)] \qquad (33)$$

where k_B is the Boltzman, T is the absolute temperature of the system, $\rho(r)$ is the number density of the protein-ligand atom pair at distance r in the training set, and $\rho^*(r)$ is the pair density in a reference state where the interatomic interactions are zero.

Posterior to the determination of the potential parameters $\mathfrak{w}(r)$, the energy of ligand binding for a given molecular complex is simply the sum of the interaction components for all the protein-ligand atom pairs in the molecular complex. Their pairwise character enables these functions to be as fast as empirical scoring functions (HUANG & ZOU, 2010).

Goals

This thesis has two distinct general aims:

1. Study the structural and dynamics effects of CSF-1R induced by D802V mutation and compare the results with those obtained for KIT in the native wild-type and mutated forms;
2. *In silico* study of the imatinib recognition by targets, where the targets consist of CSF-1R and KIT in their native wild-type and mutated forms. The mutated forms contain the oncogenic mutations D802V, in CSF-1R; D816V, V560G and S628N, in KIT. We aim to correlate the theoretical predictions with the available experimental (*in vivo* and *in vitro*) data;

In order to achieve these objectives, the specific tasks will be performed:

Objective 1 – Detailed study of the structural and dynamical features of native and mutated CSF-1R receptor

a. Generation, by comparative modeling, of CSF-1R^{WT} and CSF-1R^{D802V} mutant structures;
b. Running of molecular dynamics (MD) simulations of both receptor forms in order to obtain stable system conformations and collect the statistically significant data for analysis of their dynamical behavior;
c. PCA and Normal modes analysis to investigate the high amplitude movements in the native protein and their changes in consequence of the mutation;
d. Computation of the free energy of binding between the JMR and the TK domain in both forms of CSF-1R;
e. Comparison of the molecular network pathways between native and mutated CSF-1R.

Objective 2 – Comparative analysis of imatinib binding two different receptors from type III RTKs family: CSF-1R and KIT

a. Convergence analysis of MD trajectories of KIT (previously calculated) and CSF-1R in order to select unique conformations of wild-type and mutated CSF-1R and KIT;
b. Molecular docking of imatinib into the targets: CSF-1R and KIT in wild-type and mutated states;
c. Running of MD simulations for the imatinib+target complexes;
d. Description of imatinib binding mode in terms of H-bonds involving key conserved residues, which are described as crucial for imatinib's recognition/binding; characterization of intermolecular contacts and free energy of binding.

Chapter 2: Methodology

1. Structural and dynamical study of wild-type and mutant forms of CSF-1R

This part of the work has been published and can be also found at (DA SILVA FIGUEIREDO CELESTINO GOMES *et al.*, 2014).

1.1. Secondary structure prediction of the JMR

The secondary structure prediction was performed for the JMR residues of the wild-type (WT) CSF-1R receptor. Six methods based on the protein primary sequence were used: GOR4 (GARNIER; GIBRAT & ROBSON, 1996), Jpred (COLE; BARBER & BARTON, 2008), SOPMA (GEOURJON & DELÉAGE, 1995), SCRATCH (CHENG *et al.*, 2005), NetSurfP (PETERSEN *et al.*, 2009) and Psipred (MCGUFFIN; BRYSON & JONES, 2000).

For comparison with the sequence-based prediction methods for the secondary structure, we have additionally used STRIDE (FRISHMAN & ARGOS, 1995), a knowledge based algorithm that assigns the secondary structure based on the atomic coordinates, with a combined use of hydrogen bond energy and statistically derived backbone torsion angle information. The crystallographic structure of the wild-type auto-inhibited form of CSF-1R (PDB ID: 20GV) (WALTER *et al.*, 2007) was used as input for STRIDE.

1.2. Electrostatic potential surface

Electrostatic potential surfaces were calculated for the crystal structures of the auto-inhibited inactive form of the cytoplasmic region of CSF-1R (PDB ID: 2OGV) and KIT (PDB ID: 1T45). The software APBS (BAKER *et al.*, 2001) was used through the PDB2PQR web-based server (DOLINSKY *et al.*, 2007).

1.3. Preparation of initial coordinates

The crystallographic structure of the WT auto-inhibited inactive form of CSF-1R (PDB ID: 20GV) was retrieved from the Protein Data Bank (BERMAN, 2000). MODELLER 9v8 (ESWAR *et*

al., 2008) was used to add missing atoms at some residues (543-545, 606-607, 620-621, 623, 625, 677, 741, 812, 814 and 918). *In silico* substitution of Asp (D) to Val (V) at position 802 was also performed by MODELLER, using the WT structure (PDB ID: 2OGV) as template, making them comparable starting models. Generated models of the native CSF-1R and its mutant D802V will be referred to as CSF-1R^{WT} and CSF-1R^{MU} respectively.

1.4. Molecular dynamics simulations

Setup of the systems

The setup of the systems (CSF-1R^{WT} and CSF-1R^{MU}) was performed using AMBER force field, parameter set 99SB (HORNAK *et al.*, 2006) inside GROMACS package, version 4.5 (VAN DER SPOEL *et al.*, 2005). The molecules were centered in a cubic box with a 1.5 nm distance to the faces, under periodic boundary conditions and solvated with explicit TIP3P model water molecules (JORGENSEN & JENSON, 1998). Cl- counter ions were added when necessary to neutralize the overall charge (3 for CSF-1R^{WT} and 4 for CSF-1R^{MU}). The minimization procedure consisted of 2 steps: steepest descent energy minimization (EM) with the solute atoms restrained; (ii) EM with all atoms free. The equilibration procedure was performed on the solvent, keeping the solute heavy atoms restrained for 500 ps at 310 K and a constant volume (canonical NVT ensemble).

Production of trajectories

Two production runs of 50 ns, using different seeds for velocity generation, were carried out for both receptors, CSF-1R^{WT} and CSF-1R^{MU}. The temperatures of solute (protein) and solvent (water and ions) were separately coupled to the velocity rescale thermostat (BUSSI; DONADIO & PARRINELLO, 2007) at 310 K with relaxation time of 0.1 ps. The pressure was maintained at 1 atm by isotropic coordinate scaling with relaxation time of 1 ps using Berendsen thermostat (BERENDSEN *et al.*, 1984). A time step of 2 fs was used to integrate the equations of motion based on the Leap-Frog algorithm (VAN GUNSTEREN & BERENDSEN, 1988b). The Lennard-Jones interactions were shifted to a cut-off 1.4 nm, and the Particle Mesh Ewald (PME) method (DARDEN; YORK & PEDERSEN, 1993) was used to treat long-range electrostatic interactions. The neighbor list for the electrostatic interactions was updated every 5 steps, together with the pair list. All bonds were constrained using the P-LINCS

algorithm (HESS, 2008). The SETTLE algorithm (MIYAMOTO & KOLLMAN, 1992) was used to constrain the geometry of the water molecules. Coordinates files were recorded every 1 ps.

Analysis of the trajectories

The trajectories for each pair of molecular dynamics (MD) simulations were analyzed with tools included in the GROMACS package. When concatenating the MD simulations replicas, the first 5 ns of each replica trajectory, needed to achieve relaxation, were not considered. Analyses were performed on the resulting merged trajectory of 90 ns for each protein or based on the 45 ns individual replicas. We have also produced a 60 ns concatenated trajectory from the last 30 ns from each replica to be further used for *IDSs* calculations with MONETA (LAINE; AUCLAIR & TCHERTANOV, 2012).

A convergence analysis was performed on the merged trajectories of 90 ns using an ensemble-based approach (LYMAN & ZUCKERMAN, 2006). The algorithm was described earlier in the Introduction. The merged trajectory was split in four halves (two halves for each replica) and conformations from each half were grouped based on their RMSD from each reference structure. A good convergence quality was assessed when each reference group was populated by conformations from the four halves of the trajectory at equivalent levels, meaning that every reference structure is equivalently represented in both replicas of the trajectory.

Geometrical measurements

The module *g_dist* available in GROMACS was used to measure two characteristic distances monitored every 10 ps over the MD simulations of each model: (i) the distance **d1** between the centroid (C) of the JM-B region (residues 543-552, C1) and the C of the remaining residues in the N-lobe (582-664, C1'); (ii) the distance **d2** between the C of the JM-S (residues 553-564, C2) and the C-lobe (residues 671-922, C2').

The hydrogen (H-) bond analyses were done with the program *g_hbond* available in GROMACS. Time occupancy of H-bonds stabilizing the JMR and the A-loop was recorded every 100 ps of simulation for each model of CSF-1R. H-bonds (D•••H–A) where defined with a DHA angle cutoff of 120° and a D•••A distance cutoff of 3.5 Å (D and A are donor and acceptor atoms).

Secondary structure prevalence

The monitoring of the secondary structure content over the MD simulations was calculated using the module *do_dssp* available in GROMACS. The program makes use of DSSP (KABSCH & SANDER, 1983). The calculation was performed over the merged 90 ns trajectories for both forms of the receptor, WT and MU.

1.5. Energy analysis

The free energy of JMR or its segments (ligand, L) binding to KD (receptor, R) defined as

$$\Delta G_{bind} = G_{RL} - (G_R + G_L) \quad (35)$$

was computed over the merged trajectories and on the individual MD simulations, considering only the last 30 ns from each replica for both CSF-1R^{WT} and CSF-1R^{MU}.

Free energies were evaluated using the Molecular Mechanics Generalized Born Surface Area (MM-GBSA) method, implemented in AMBER 12 (BASHFORD & CASE, 2000; KOLLMAN *et al.*, 2000; ONUFRIEV; BASHFORD & CASE, 2000, 2004).

This method, explained also in the Introduction, combines the molecular mechanical energies with the continuum solvent approaches. The molecular mechanical energies represent the internal energy (covalent bonds, angles and dihedral angels contributions), and contribution of van der Waals and electrostatic interactions. The electrostatic contribution to the solvation free energy is calculated by generalized Born (GB) methods (KOLLMAN *et al.*, 2000). The non-polar contribution to the solvation free energy is determined with solvent-accessible-surface-area-dependent terms. Estimates of conformational entropies are calculated with the normal mode module from AMBER.

1.6. Normal modes analysis (NMA)

NMA was performed using the diagonalization in a mixed basis (DIMB) method (PERAHIA & MOUAWAD, 1995) of the VIBRAN module of CHARMM 35b3 (BROOKS *et al.*, 1983, 2009) on MD conformations from (i) CSF-1R^{WT} taken at 1.526, 49.390, 66.530 and 81.680 ps, spanning both replicas contained in the 90-ns merged trajectory, and (ii) CSF-1R^{MU} mutant taken at

5.510, 23.530, 40.670 and 84.680 ps. The selected MD conformations were found to be the most representative of the trajectories, according to the convergence analysis.

The first hydration shell (5 Å) around the MD conformations was kept to help prevent the solvent-exposed regions of the protein from collapsing during the minimization procedure (BATISTA *et al.*, 2010). During initial steepest descent energy minimization of the system, mass-weighted harmonic constraints of 250 kcal/mol/A^2 were applied to the starting structure and reduced by a factor of 2 every 1000 minimization steps until they fell below a threshold value of 5 kcal/mol/A^2. The constraints were then removed and the system was minimized by conjugate gradient and adopted-basis Newton-Raphson steps until the RMS energy gradient fell below 10^{-5} kcal/mol/A^2. Normal modes were computed by diagonalizing the mass-weighted Hessian matrix of the energy-minimized conformations and the 96 non-zero lowest-frequency modes were analyzed.

The degree of collectivity of the JMR motions in a given mode *l* was calculated as (BRÜSCHWEILER, 1995; TAMA & SANEJOUAND, 2001):

$$k_{JMR}(l) = \frac{1}{n}\exp(-\sum_{i=1}^{n} \alpha_i(l)^2 \ln\left(\alpha_i(l)^2\right)) \qquad (36)$$

where n=663 is the number of atoms belonging to the JMR. The quantity α_i is defined as:

$$\alpha_i(l) = \frac{x_i(l)^2 + y_i(l)^2 + z_i(l)^2}{\sum_j^n x_i(l)^2 + y_i(l)^2 + z_i(l)^2} \qquad (37)$$

where x_i, y_i and z_i are the components of mode l showing the three degrees of freedom of atom i and such that $\sum_i^n a_i^2 = 1$.

The degree of collectivity is comprised between 0 and 1. A value of $1/n$ indicates that only one atom is involved in the motion while a value close to 1 indicates high collectivity. The resultant displacement, *i.e.* the norm of the resultant displacement vector, of any fragment of the protein was calculated as:

$$R = \sqrt{(\sum_i^m x_i)^2 + (\sum_i^m y_i)^2 + (\sum_i^m z_i)^2} \qquad (38)$$

over the ensemble M of the *m* atoms belonging to the fragment –172 for JM-Switch and 181 for JM-Zipper.

1.7. Principal Component Analysis (PCA)

PCA was applied to each model to identify the main eigenvectors (3N directions) along which the majority of the collective motions are defined. The calculations were performed on the backbone atoms positions recorded every ps along the trajectories for each 45 ns simulation replica. The 100 first modes of each trajectory were extracted. The calculation was performed using the *g_covar* module of GROMACS package.

The overlap between the first 10 modes of each trajectory was calculated using the *g_anaeig* module of GROMACS package. Briefly, the method consists in overlapping the subspace spanned by ***m*** orthogonal vectors $\mathbf{w_1},\ldots,\mathbf{w_m}$ with a reference subspace spanned by *n* orthonormal vectors $\mathbf{v_1}, \ldots,\mathbf{v_n}$ and it can be quantified as follows:

$$overlap(v,w) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}(v_i.w_j)^2 \qquad (39)$$

The overlap will increase with increasing ***m*** and will be 1 when set **v** is a subspace of set **w**.

1.8. Analysis of intramolecular communication

Modular network representations of CSF-1R^{WT} and CSF-1R^{MU} were built and visualized with MONETA (LAINE; AUCLAIR & TCHERTANOV, 2012), using the most advanced version (ALLAIN *et al.*, 2014). *IDSs* were identified from Local Feature Analysis (LFA) (PENEV & ATICK, 1996) based on PCA. PCA calculations were performed for both models of the receptor, on the Cα atoms covariance matrices calculated on the concatenated 60 ns trajectory merged from the two 50 ns MD replicas, considering only the last 30 ns of each simulation. From the *3N* eigenvalues associated with the *3N* eigenvectors, the first 17 and 19 eigenvectors were sufficient to describe 80% of the total Cα atomic fluctuations on CSF-1R^{WT} and CSF-1R^{MU}, respectively. These vectors were used to apply the LFA formalism. A threshold value P_{cut} was arbitrary chosen by the program to keep 1.0 % of all LFA cross-correlations above it. The value was set to 0.035 for the WT and 0.038 for the D802V CSF-1R.

Distance matrices consisting of the average of the smallest distance between each residue pairs were computed using the *g_mdmat* module of GROMACS package, v.4.5.6, considering only the C-α atoms. Two residues were considered neighbors if the average smallest distance between them was lower than a given threshold d_{cut} of 3.6 Å. Since we have observed a slightly different dynamical behavior in the two MD simulation replicas, we have computed the *CPs* on the individual MD simulations, considering the last 30 ns only, in order to distinguish between the communication pathways of CSF-1R^{WT} and CSF-1R^{MU}. One replica of each form of receptor was retained for the illustrations.

The *CPs* are grown ensuring that the adjacent residues are connected by non-covalent interactions and that every residue in the *CP* is connected to any other point by a shorter commute time (*CT*). Non-bonded interactions were recorded along the MD simulations using LIGPLOT (WALLACE; LASKOWSKI & THORNTON, 1995). Two residues were considered as interacting when they formed at least one non-bonded interaction for 50% of the simulation time. To discriminate between large and short *CT*s, a threshold CT_{cut} was chosen so that highly connected residues communicate efficiently with about 10% of the total number of residues in the protein (CHENNUBHOTLA & BAHAR, 2007). The threshold values were set to 0.1for both models.

Statistical analyses were performed with the R software (IHAKA & GENTLEMAN, 1996); visualization of the structure/interaction/communication characteristics/results are performed with PyMOL (DELANO, 2004) incorporated in MONETA (ALLAIN *et al.*, 2014).

2. CSF-1R and KIT receptors complexed with imatinib

2.1. Preparation of initial coordinates

We decided to use previous equilibrated conformations of CSF-1R and KIT as starting models for docking simulations. The structures of CSF-1R and KIT in their WT and mutant forms (CSF-1R^{D802V}, KIT^{V560G}, KIT^{S628N} and KIT^{D816V}) were retained from MD simulations replicas run using the same parameters and force field (AMBER parameter set 99sb). For CSF-1R, the trajectories computed in the Part I were used (DA SILVA FIGUEIREDO CELESTINO GOMES *et al.*, 2014); for KIT we have collected the trajectories computed previously (CHAUVOT DE BEAUCHÊNE, 2013; CHAUVOT DE BEAUCHÊNE *et al.*, 2014; VITA *et al.*, 2014).

After retrieving all the trajectories, we were interested in finding "unique" representative conformations of CSF-1R and KIT that were sampled only at the WT or at the mutant trajectories. In order to search these conformations, we have re-written the MD trajectories taking into account only the Cα atoms so that we could combine the WT and mutant trajectories in only one concatenated trajectory. This `merge` was done separately for each receptor.

Considering that each one of the two replicas for each receptor form contains 5000 frames, we discarded the first 5ns from the calculation (first 500 frames). The resulting combined trajectory contained 18000 frames for CSF-1R (WT and MU) and 36000 frames for KIT (WT, V560G, S628N and D816V). Further, we applied the convergence analysis (LYMAN & ZUCKERMAN, 2006). In this particular case, we were not interested in finding the conformations best represented in all the ensemble of frames, but the structures sampled only at the regions of the trajectory corresponding to the WT or the mutants' conformational pool of frames. Using a RMSD cut-off **r** of 2.0 Å, we were able to find a few unique reference structures for the WT CSF-1R and KIT, and also for the mutants.

JMR residues were not taken into account at the moment of convergence since they are very flexible. Early attempts to dock imatinib into the inactive structures of KIT, either WT or mutant, have failed (CHAUVOT DE BEAUCHÊNE, 2013) due to the buried conformation of the JMR, with its N-ter inserted into the ATP-binding site. Therefore, before docking, we have decided to exclude the JMR from the TK domain in all systems (residues 543-581 in CSF-1R; residues 547-588 in KIT), with exception of the mutant KIT^{V560G}, in which the mutation site is found in the JMR. For this mutant, we truncated the JMR in position 558.

A possible steric hindrance was detected by the superposition of the auto-inhibited structures of CSF-1R (PDB ID: 2OGV) and KIT (PDB ID: 1T45) with the structure of inactive KIT complexed with imatinib (PDB ID: 1T46). The phenylalanine F811 in the A-loop DFG motif is also inserted into the ATP-binding site in the auto-inhibited structure of KIT, such impairing the inhibitor binding. The same observation is valuable in crystal structure 2OGV. Consequently, among the few "unique" structures issued from the convergence analysis of merged trajectories, we have chosen the ones that presented the absence of possible steric clashes with imatinib. Criteria of selection was a superimposing of conformations-candidates with the 1T46 crystal structure.

RTKs are ATP-dependent phosphotransferases that deliver a single phosphoryl group from the γ position of ATP to the hydroxyls groups of tyrosine in protein substrates. They require an essential divalent metal ion, usually Mg^{2+}, to facilitate the phosphoryl transfer reaction and assist in ATP binding (ADAMS, 2001).

Since we are not dealing with ATP, the presence of metals inside the ATP-binding site were not considered. KIT receptor complexed with imatinib (PDB ID: 1T46) (MOL *et al.*, 2004) did not contain Mg^{2+} ions, the same with KIT or CSF-1R complexed with another inhibitors (PDB IDs: 3LCO, 3LCD, 4HW7, 2I1M, 2I0Y, 3DPK, 2I0V, 3BEA, 4U0I, 3G0E, 3G0F, 4HVS). In addition, the presence of cationic species in the binding site could possibly compromise the docking of imatinib, since the ligand is protonated on the docking and MD simulations.

2.2. Molecular docking

The 3D structure of imatinib was retrieved from the PDB file of KIT complexed with the inhibitor (PDB ID: 1T46). The structures preparation of the receptors and the inhibitor imatinib, as well as the docking runs, were performed using the Schrödinger suite Maestro ("Schrödinger Release 2014-2: Maestro", 2014). The *protein preparation wizard* was used to re-assign hydrogens, charges and to minimize the structures of WT CSF-1R/KIT and mutants, using the default parameters.

Imatinib was prepared using *LigPrep* ("Schrödinger Release 2014-2: LigPrep", 2014) in the environment at pH 7.0 with a pH threshold of 2 points. From the possible ligand protonation states that were generated by the program, we have chosen the protonated form of imatinib (+1), which seems to be the correct protonation state for the inhibitor in complex with kinases, according to a previous study (ALEKSANDROV & SIMONSON, 2010).

The receptor structures of CSF-1R and KIT WT and mutated forms (one for each), selected as explained in section 2.1, were superposed to the crystal structure of KIT complexed with imatinib (PDB ID: 1T46) in order to center the docking workspace on the ligand. All residues that were within a 5 Å cut-off of imatinib were allowed to be flexible (~30 residues). Docking simulations were performed using the *InducedFit* (IFD) protocol (SHERMAN *et al.*, 2006) with the *Extended sampling*, which performs automated, extended sampling in the initial stages with optimized docking settings. Complexes were chosen based on the docking energy (Glide

and IFD score) and the RMSD values in relation to Imatinib co-crystallized with KIT (PDB ID: 1T46). The validation of our docking protocol was knowledge-based since the docking resulting poses were very similar to the imatinib's conformation in the structure of KIT complexed with the inhibitor (1T46).

The attempts of docking imatinib into the mutant KIT^{V560G} had poor energy conformations, in comparison with the other systems using similar parameters, mentioned above. Therefore, we have defined a constraint that restricts the docking of imatinib to the target within a specified RMSD tolerance of the core of a reference ligand. As the reference, we used the crystallographic structure of inactive KIT complexed with imatinib (PDB ID: 1T46). RMSD tolerance was set to 2.0 Å.

2.3. Molecular dynamics simulations of the imatinib-target complexes

Imatinib parametrization and topology construction

The structure of imatinib was retrieved from the Zinc Database in the mol2 format. The topology to be used as input for the MD simulations was generated using the web-based server Swissparam (ZOETE *et al.*, 2011), which generates topology parameters compatible with CHARMM all-atoms force field for use in GROMACS package (VAN DER SPOEL *et al.*, 2005). The protonation state of imatinib and the charges of the automated generated topology were replaced by the parameters rigorously defined by an earlier study from Aleksandrov *et al.* (ALEKSANDROV & SIMONSON, 2010), where the authors simulated the possible protonation states of the imatinib in complex with the Abl receptor by accurate MD simulations using CHARMM. The results indicated that imatinib binds to Abl in its protonated, positively charged form (+1). The topology file used in the simulations can be found in section II of the Appendix.

Setup of the systems

The setup of the generated complexes was performed using CHARMM27 all-atom force field (MACKERELL; BANAVALI & FOLOPPE, 2000) integrated in GROMACS package (VAN DER SPOEL *et al.*, 2005), version 4.6.5. The complexes will be referred as $CSF\text{-}1R^{WT}$, $CSF\text{-}1R^{D802V}$, KIT^{WT}, KIT^{V560G}, KIT^{S628N} and KIT^{D816V}. In addition, we have performed simulations for the WT

form of CSF-1R and KIT in absence of any ligand and without the JMR domain. They will be referred as CSF-1R^{apo} and KIT^{apo}.

The remaining setup was done as for the previous MD simulations of CSF-1R^{WT} and CSF-1R^{MU}. Namely, each system was centered in a cubic box with a 1.2 nm distance to the faces, under periodic boundary conditions and solvated with explicit TIP3P model water molecules (JORGENSEN & JENSON, 1998); Cl^- counter ions were added when necessary to neutralize the overall charge. The minimization procedure consisted of 2 steps: steepest descent energy minimization (EM) with the solute atoms restrained; (ii) EM with all atoms free. The equilibration procedure was performed on the solvent, keeping the solute heavy atoms restrained for 500 ps at 310 K and a constant volume (canonical NVT ensemble).

Production of the trajectories

Two production runs of 50 ns were carried out for all complexes and for the apo structures. The temperatures of the solute (receptor + imatinib) and solvent (water and ions) were separately coupled to the velocity rescale thermostat (BUSSI; DONADIO & PARRINELLO, 2007) at 310 K with relaxation time of 0.1 ps.

The remaining parameters were the same as for the previous MD simulations for CSF-1R: the pressure was maintained at 1 atm by isotropic coordinate scaling with relaxation time of 1 ps using Berendsen thermostat (BERENDSEN *et al.*, 1984); a time step of 2 fs was used to integrate the equations of motion based on the Leap-Frog algorithm (VAN GUNSTEREN & BERENDSEN, 1988b); the Lennard-Jones interactions were shifted to a cut-off 1.4 nm, and the Particle Mesh Ewald (PME) method (DARDEN; YORK & PEDERSEN, 1993) was used to treat long-range electrostatic interactions; the neighbor list for the electrostatic interactions was updated every 5 steps, together with the pair list; all bonds were constrained using the P-LINCS algorithm (HESS, 2008); and the SETTLE algorithm (MIYAMOTO & KOLLMAN, 1992) was used to constrain the geometry of the water molecules. Coordinates files were recorded every 1 ps.

Analysis of the trajectories

The trajectories for each pair of molecular dynamics (MD) simulations (two replicas) were analyzed with the tools included in the GROMACS package. A merged trajectory containing

9000 frames was generated from the two individual MD replicas, discarding the first 5 ns from each simulation.

The hydrogen-bond occurrences were calculated using the *g_hbond* module from GROMACS. The analysis was performed over the merged trajectories and the occurrences are represented by a percentage (%) over the simulation time.

Visual inspection of the trajectories were done with PyMOL (DELANO & LAM, 2005) and VMD (HUMPHREY; DALKE & SCHULTEN, 1996). Graphs were generated using Grace (http://plasma-gate.weizmann.ac.il/Grace/) and SciDAVis (http://scidavis.sourceforge.net).

Electrostatic surface calculation

The electrostatic surface of the protein was performed with the APBS software (LEE; DUAN & KOLLMAN, 2000) through the PDB2PQR webserver (DOLINSKY *et al.*, 2007). The structures used corresponded to the equilibrated complexes before the start of the MD simulations replicas.

2.4. Energy analysis

The free energy of binding between the ligand (L), in this case imatinib, and the receptors (R) in their WT and mutant forms defined as in equation 34 was computed over the concatenated 90ns trajectories. The free energies of binding were evaluated using the Molecular Mechanics - Poisson Boltzmann Surface Area (MM-PBSA) method, already described at the Introduction.

The ΔG analysis was performed through the recently developed *g_mmpbsa* module (KUMARI *et al.*, 2014), adapted for using with GROMACS. The tool combine subroutines from GROMACS and APBS (BAKER *et al.*, 2001) to calculate the enthalpic components of MM-PBSA interaction. Since we are only interested in the relative order of binding affinities, the entropic contribution was omitted to avoid unnecessary computational time. This is possible because we have used the same compound with proteins of similar structures and same initial binding mode. As output data, we obtain the relative binding energy for the complexes and also the contribution of each protein residue for the binding energy.

2.5. MD simulations for KIT forms containing the truncated JMR

In order to investigate if the JMR's role in the binding energy of imatinib into the targets, we have decided to perform MD simulations for the KIT forms (KIT^{WT}, KIT^{S628N} and KIT^{D816V}) containing the same truncated portion present at KIT^{V560G}.

We have taken the structures of KIT^{WT}, KIT^{S628N} and KIT^{D816V}, derived from the convergence analysis described previously, excising the portion corresponding to residues 547-558, making all KIT models equivalent in size. The truncated KIT structures were superposed with the final docked structures containing the low energy conformation of imatinib and the inhibitor was placed manually into the ATP-binding site of the truncated structures. The possible steric clashes were eliminated by minimizing the obtained complexes using the same protocol as described in section 2.3 of this chapter. Equilibration and MD simulations (two replicas of 50ns) were performed under the same conditions as for the previous simulations.

Chapter 3: Results

1. Structural and dynamic study of the wild-type CSF-1R and the D802V mutant: comparison with the effects observed in KIT^{D816V}

The models of the native cytoplasmic region of CSF-1R (residues 543-922) and its mutant containing the D802V substitution (referred here as CSF-1R^{WT} and CSF-1R^{MU} respectively) were generated from the crystallographic structure of the wild-type (WT) receptor in its auto-inhibited inactive state (PDB ID: 20GV) (WALTER *et al.*, 2007). A similar abbreviation for KIT will be used in the text for cross-receptor comparison: KIT^{WT} and KIT^{D816V}, the latter being the homologous mutation in KIT.

Molecular dynamics (MD) simulations of the generated models (two 50 ns trajectories for each form) were carried out to investigate and compare the structure and internal dynamics of the two proteins, CSF-1R^{WT} and CSF-1R^{MU}. The global dynamical behavior of the proteins was explored by measuring the root mean square deviations (RMSDs) of backbone atoms with respect to the initial frame plotted versus simulation time and showed separately for N- and C-lobes, the JMR and the A-loop regions (Fig. 17 A).

The four trajectories of CSF-1R (two replicas for CSF-1R^{WT} and two for CSF-1R^{MU}) displayed comparable conformational drifts, with RMSD mean values in the range 0.12 – 0.30 nm indicating a tolerable stability of the simulated systems after a 5 ns relaxation interval. However, the RMSD profile of the A-loop region showed a high deviation after 17 ns for one replica of CSF-1R^{MU}, with RMSD values up to ~0.26 nm, which was not observed in the other trajectory replicas. We observed a similar behavior for the A-loop in KIT MD simulations (LAINE *et al.*, 2011), although the deviations were significantly larger than in CSF-1R (reaching 0.46 nm in KIT).

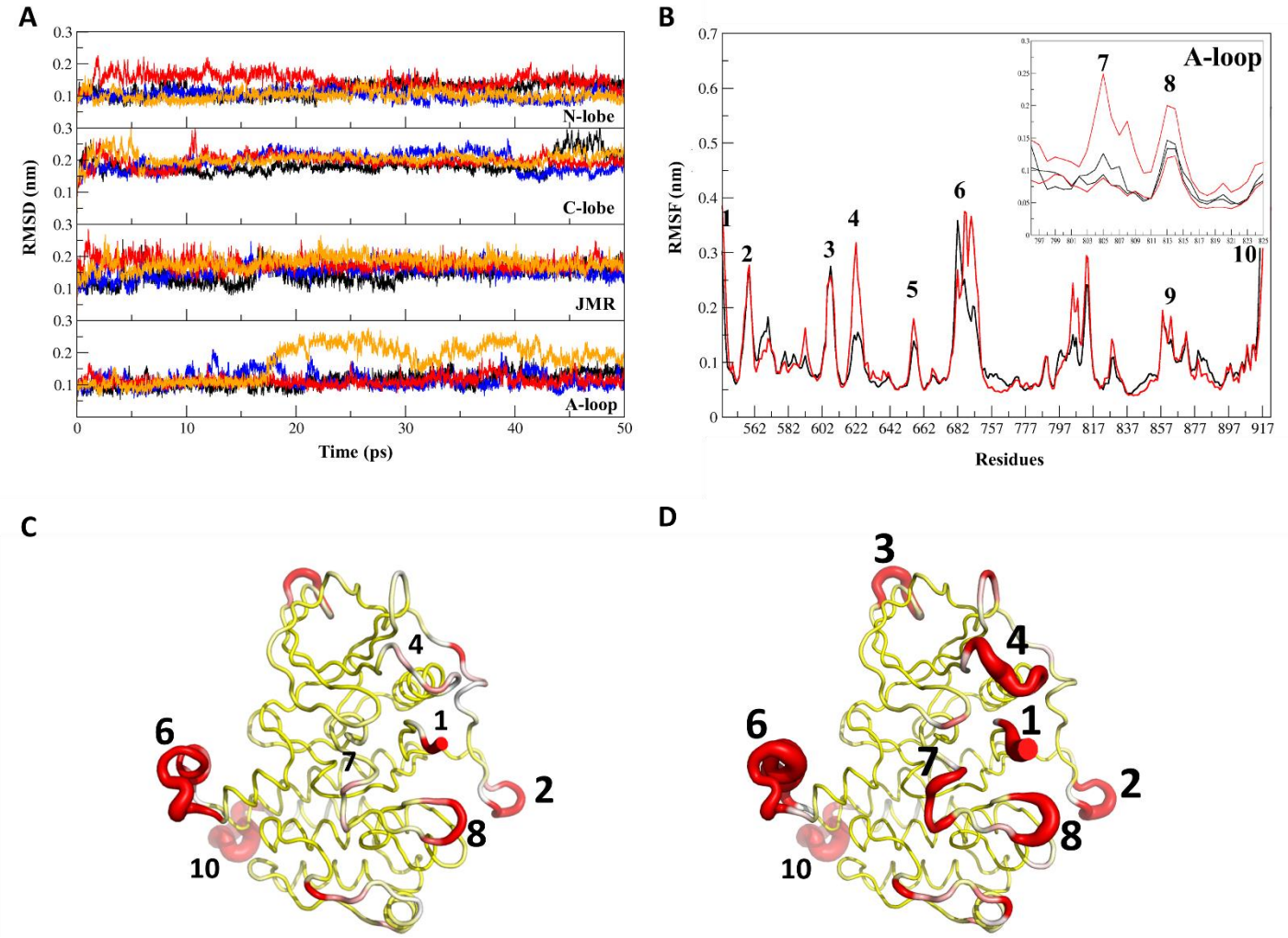


Figure 17: ***MD simulation data for CSF-1R inactive form****. Two forms of receptor, the native (CSF-1R^{WT} and CSF-1R^{MU} (D802V) were simulated twice during 50 ns. (**A**) The Root Mean Square Deviation (RMSD) values were calculated for backbone atoms from trajectories 1 and 2 of MD simulations of CSF-1R^{WT} (black and blue) and CSF-1R ^{MU} (red and orange). RMSDs (in nm) plotted versus simulation time (ns) and showed separately for N- and C-lobes, JMR and A-loop regions. (**B**) The Root Mean Square Fluctuations (RMSF) computed on the backbone atoms over the total production simulation time of CSF-1R^{MU} (red) were compared to those in CSF-1R^{WT} (black). The RMSFs of the A-loop is zoomed in the insert. The average conformations for CSF-1R^{WT} (**C**) and CSF-1R^{MU} (**D**) are presented as tubes. The size of the tube is proportional to the by-residue atomic fluctuations computed on the backbone atoms. The high fluctuation region found in proteins, are specified by red color and numerated from 1 to 10 in B–D. The size of numbers in **D** is proportional to RMSFs.* (DA SILVA FIGUEIREDO CELESTINO GOMES *et al.*, 2014)

The root mean square fluctuation (RMSF) values, describing atomic fluctuations averaged over the protein residues, ranged from 0.1 to 0.4 nm, and were overall quantitatively comparable between CSF-1R^{WT} and CSF-1R^{MU} (Fig. 17B). Projection of RMSF values on the tridimensional structure of CSF-1R (Figs. 17C and 17D) revealed that the most flexible residues formed clusters located in the JMR, encompassing the most buried JM-B fragment (residues 543-545) and part of the JM-S (residues 556-560), the A-loop, the KID, and the loop that connects β3-strand (residues 620-625) and Cα-helix in the N-lobe.

The D802V mutation noticeably enhanced RMSF fluctuations in these regions (Fig. 17). A zooming on the A-loop RMSF values indicated the perturbation on the atomic coordinates observed in one of the MD simulations of CSF-1R^{MU} (Fig. 17, insert).

The crystallographic data of the native receptor (PDB ID: 2OGV) (WALTER *et al.*, 2007) shows that residue D802 is located in a short bend between two small 3_{10}-helices *(H)* formed by residues 798-800 (*H1*) and 803-805 (*H2*). By analyzing the crystal structures of auto-inhibited KIT and FLT3, it was suggested that the side chain of Asp at this position (D835 in FLT3 and D816 in KIT) contributes to stabilize the inactive kinase structure, perhaps through hydrogen bonding with nearby residues in the P-loop (GRIFFITH *et al.*, 2004; MOL *et al.*, 2004).

In addition, the negatively charged side chain of Asp might act to stabilize the charge distribution (dipole moment) of the adjacent helix, *H2*. The side chain of D816 in KIT is positioned in such a way that it could interact with the positive end of the helix dipole formed by residues Ile817-Asp820, which would be expected to stabilize its structure and, in turn, the inactive structure of the A-loop (DIBB; DILWORTH & MOL, 2004). The Asp substitution to a Val in KIT has been proved to be potent oncogenic since it is commonly found in cancer (FLETCHER & RUBIN, 2007). The same substitution at CSF-1R has been proved to be strongly transforming in cell lines (GLOVER *et al.*, 1995) and activate the CSF-1R receptor (MORLEY *et al.*, 1999).

Systematic analysis of the MD conformations indicated that the structure of CSF-1R cytoplasmic region was globally conserved over the simulation time in CSF-1R^{WT} and CSF-1R^{MU} and shows a general similarity between these two forms (Fig. 18). Nevertheless, a detailed inspection of the secondary structures showed different folding of the A-loop in the two proteins.

Over the MD simulations of CSF-1R^{WT}, the structure of *H1* region was mainly folded as a 3_{10}-helix while the *H2* region secondary structure type alternated between 3_{10}-helix (5%), bend (20%), turn (30%) and coil (45%) (Fig. 19 B). The H2 of CSF-1R^{WT} has a different composition than in KIT^{WT}. While in KIT, the Asp in 816 position is stabilizing a Lys, in CSF-1R, there is a Met in the same position, which explains why this structure fluctuates a lot in the WT form. In CSF-1R^{MU}, the only secondary structure element retained over the simulations is the 3_{10}-helix *H1* positioned prior the D802V mutation site. The second 3_{10}-helix, *H2*, which follows the mutated site, is mainly unstructured, and the residues 803-805 adopt a turn conformation as was evidenced for most of the simulation time.

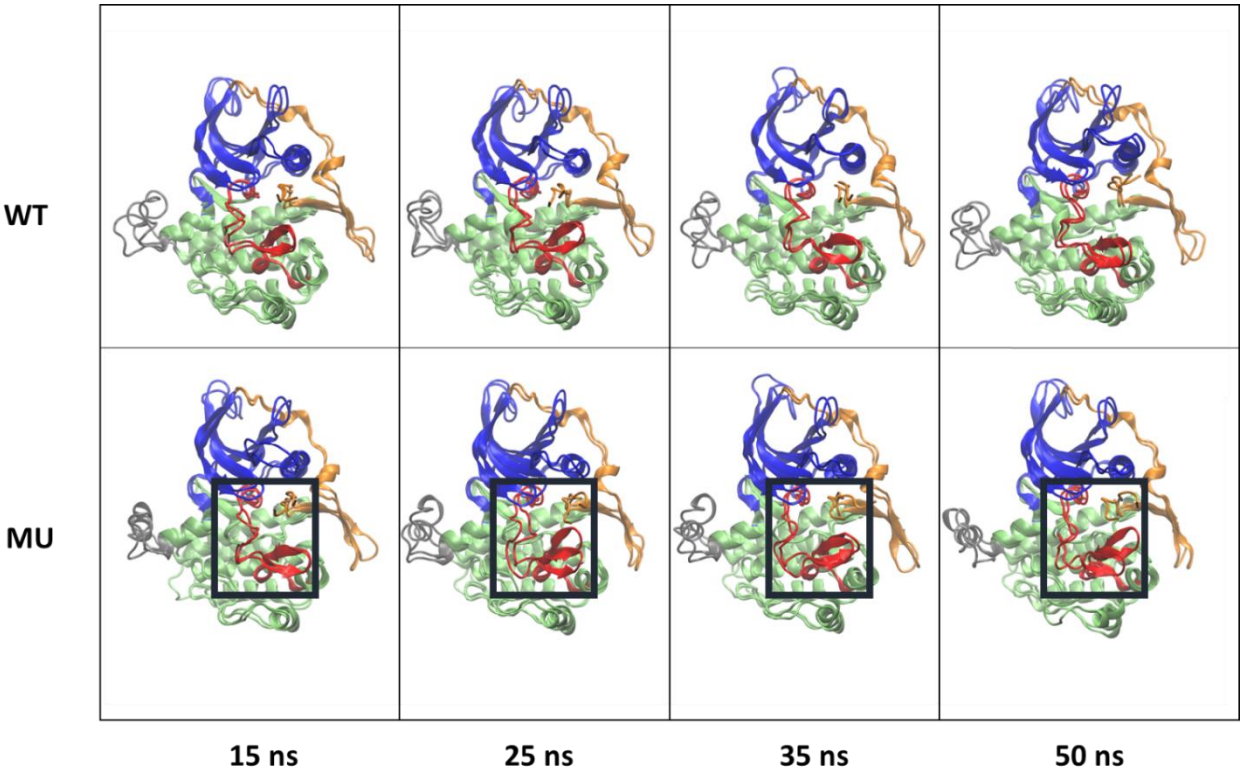


Figure 18: ***MD conformations of CSF-1R cytoplasmic region in the native protein and its D802V mutant****. Ribbon diagrams display the proteins regions or fragments with different colors: JMR (orange), A-loop (red), N- and C-lobe (blue and green), and KID (gray). Snapshots taken from the two MD replicas at 15, 25, 35 and 50 ns for CSF-1R^{WT} (top) and CSF-1R^{MU} (bottom) were superimposed by pair. Superposed conformations were selected by RMSDs clustering. The change in the A-loop conformation in CSF-1R^{MU} is highlighted with a black box.*

Such disappearing of the well-ordered structural element, previously observed in KIT^{WT}, and the increased atomic fluctuations in the A-loop, could be a result from the disruption of a positive dipole moment formed by the small 3_{10}-helix adjacent to the mutation, which is supposed to destabilize the inactive structure of the A-loop, as mentioned above. A similar

local structural effect was observed experimentally in KIT^{D816H} (PDB ID: 3G0F) (GAJIWALA *et al.*, 2009) and predicted by *in silico* studies in KIT^{D816V} (LAINE *et al.*, 2011) and in KIT$^{D816H/N/Y}$ mutants (CHAUVOT DE BEAUCHÊNE *et al.*, 2014).

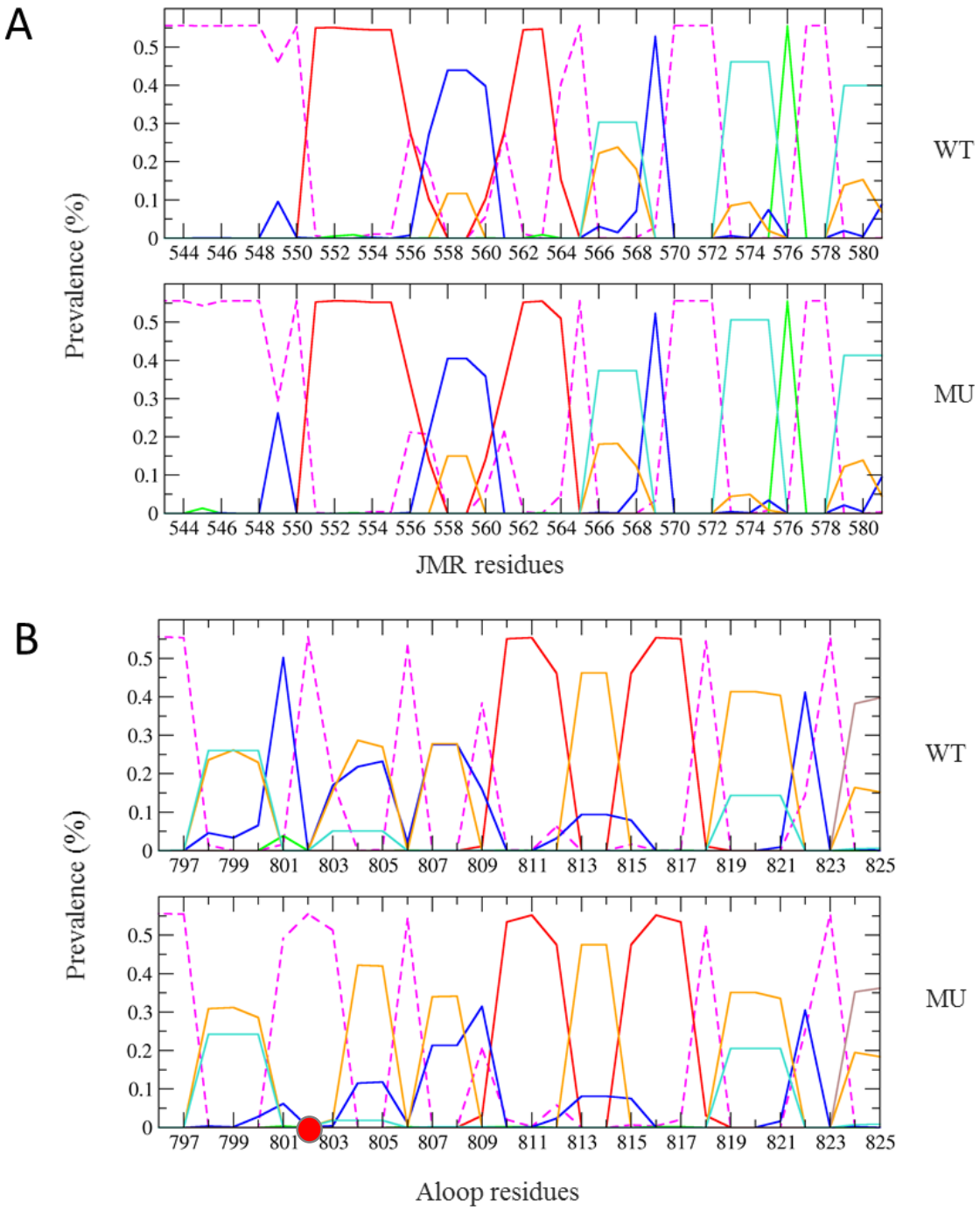


Figure 19: ***Secondary structure prevalence during MD runs.*** *Secondary structure assignments for JMR (A) and A-loop (B) were averaged over the two 50-ns MD simulations of CSF-1R^{WT} and CSF-1R^{MU}. For each residue, the proportion of every secondary structure type is given as a percentage of the total simulation time. Each secondary structure type is shown with lines of different colors: 3_{10} helices (in cyan), parallel β-sheet (in red), turn (in orange), bend (in blue) and bridge (green). Coiled structure is shown by dashed purple lines. The D802V position is indicated as a red circle.*

Whereas KIT D816V/H/N/Y mutations led systematically to a global structural reorganization of the JMR which adopts a well-shaped anti-parallel β-sheet structure translated in the axial position respectively to the TK domain (LAINE *et al.*, 2011), such a long-range structural effect, surprisingly, was not observed in CSF-1R^{MU}. The JMR structure and dynamics were strikingly similar in CSF-1R^{WT} and CSF-1R^{MU}. The quantitative analysis of the secondary structure pattern over the MD simulations revealed a retained secondary structure of the JMR in CSF-1R^{MU} compared to CSF-1R^{WT} (Fig. 18 A). Moreover, despite a topical increase of the JM-B fluctuations in CSF-1R^{MU}, the JMR position was rigorously maintained relative to the KD (Fig. 19 A). On the contrary to KIT^{WT}, the JMR of CSF-1R^{WT} is already folded as a well-shaped anti-parallel β-sheet, as evidenced in the crystallographic structure (WALTER *et al.*, 2007).

Altogether, KIT D816V and the homologous CSF-1R D802V mutations similarly affect the receptor structure at the proximity of the mutated residue, while the JMR structure is only altered in KIT mutant. This difference could be related to the distinct sequence of these regions in the two receptors. The A-loop D816 residue in KIT stabilize a lysine residue, while in CSF-1R we have a methionine in the same position of the lysine, so the D802 residue is solvent-exposed and do not make H-bonds with any other A-loop residue (not shown). Perhaps the mutation in KIT can lead to an increased unfolding of the A-loop, due to the now solvent-exposed lysine residue in the A-loop of KIT. This could contribute to a facilitated unfolding of this segment in KIT, and consequently disruption of the A-loop interactions with the TKD domain, such as the H-bonds between the A-loop Y823 and the catalytic D792, evidenced by the MONETA analysis in a previous publication (LAINE; AUCLAIR & TCHERTANOV, 2012).

To explore the secondary structure profile of CSF-1R's JMR (residues 538-580), we used six sequence-based secondary structure prediction methods and one structure knowledge-based method. The predictions indicated a relatively high probability of the polypeptide organization in well-folded structural elements, particularly β-strands in the segments 551-555 and 563-564 linked by a random coil including 4 residues, probably stabilized as a turn (Fig. 20).

This secondary structure prediction matches well with the JMR structure of the native receptor (CSF-1R^{WT}) observed by X-ray crystallography and predicted by STRIDE, that uses

information of the PDB structure. This observation prompts to hypothesize that either the JMR structure in CSF-1R does not depend on the TK domain – a behavior quite different from the allosterically regulated JMR folding in KIT, – or D802 in CSF-1R and D816 in KIT do not play a equal role in the activation mechanisms.

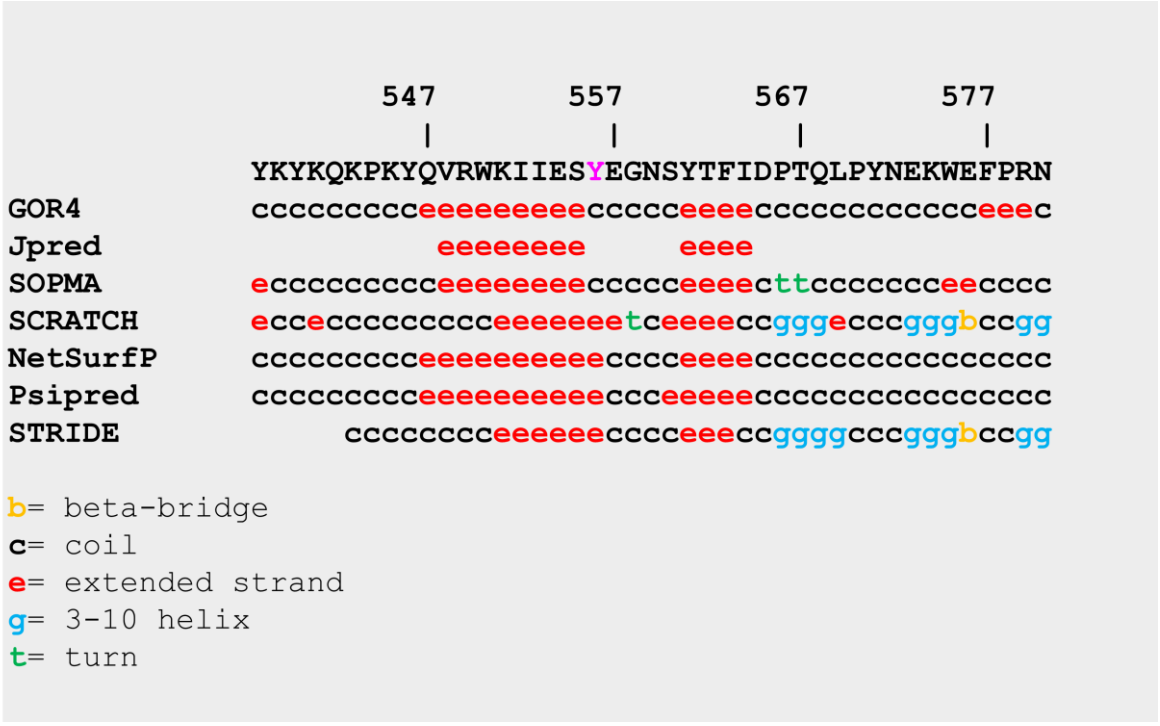


Figure 20: ***Secondary structure prediction of the JMR sequence (residues 538–580) from CSF-1R^{WT}.*** *The prediction was performed using sequence-based algorithms GOR4, Jpred, SOPMA, SCRATCH, NetSurfP, Psipred and a structure-based method STRIDE. Predicted structural elements are coded as indicated at bottom.*(DA SILVA FIGUEIREDO CELESTINO GOMES *et al.*, 2014)

In order to investigate the coupling between the JMR and the TK domain of CSF-1R, we first characterized the relative position of these two segments using two geometrical parameters, *d1* and *d2*, describing the distance between the centroids defined on JM-B and N-lobe, and JM-S and C-lobe, respectively (Fig. 21). Monitoring of these distances over the MD simulations indicated a very slight increase (~ 0.15-0.2 nm) of *d1* from the initial value observed in only one MD trajectory of CSF-1R^{MU}. The *d2* profiles of the two proteins blend into each other, demonstrating that JM-S and C-lobe retained their relative position in the mutated receptor.

Since we could not observe any signs of JMR departure by visualizing the MD simulation and monitoring the distance between this domain and the TK domain, we decided to investigate the essential dynamics of CSF-1R in both forms. The large-amplitude collective

motions generally describe the protein functional dynamics (BERENDSEN, 2000). Among these motions, the most relevant ones, also known as the softest modes, are usually highly collective, *i.e.,* they drive the cooperative motions of entire domains/ subunits.

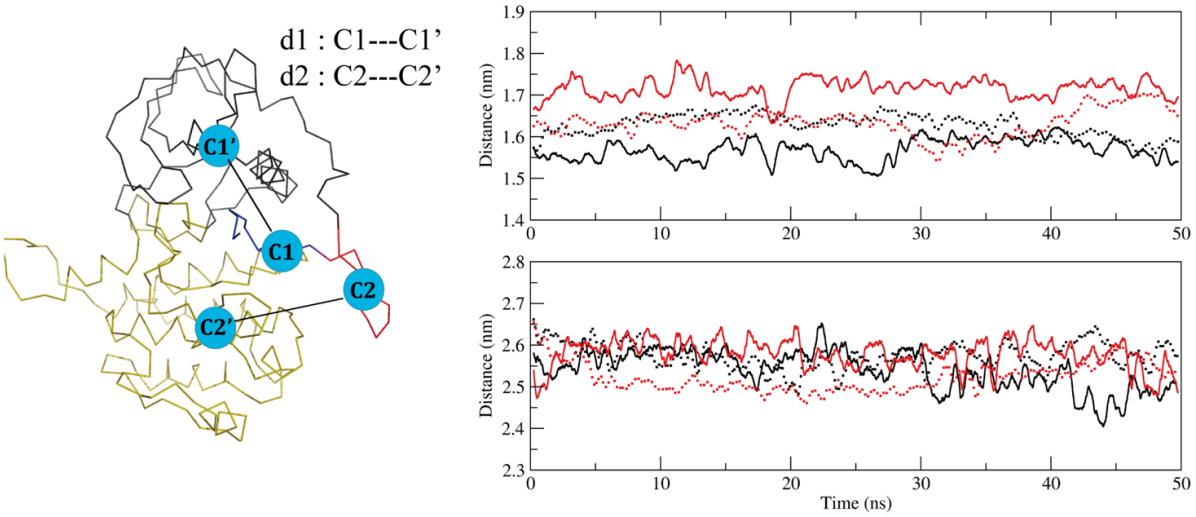


Figure 21: ***Distance monitoring between the JMR and the TK domain of CSF-1R****. Left: Distances d1 and d2 between the centroids C1 (JM-B)) and C1' (N-lobe) and between C2 (JM-S) and C2' (C-lobe), respectively. Right : Distance d1 (at the top) and d2 (at the bottom) monitored during the two replicas of the 50 ns MD simulations (full and dashed lines) for CSF-1R^{WT} (black) and CSF-1R^{MU} (red).*

We then used Principal Component Analysis (PCA) (i) to clarify the mutation induced effects in the context of collective motions between functional CSF-1R fragments in the cytoplasmic region, (ii) to compare the impact of mutation on dynamical features of CSF-1R and KIT, and (iii) to connect motions with communications between two spatially distant regulatory fragments: A-loop, Cα-helix and JMR.

The most relevant movements of CSF-1R fragments were identified by emphasizing the amplitudes (eigenvalues) and directions (eigenvectors) of the protein motions dominating the residue pair covariance matrix calculated from the MD conformational ensemble. The calculation was done for the individual MD simulation trajectories of each model and the best overlap between CSF-1R^{WT} and CSF-1R^{MU} replicas was used for illustration. Among the first 10 eigenvectors, which contribute the most to the total atomic fluctuations, the first two modes of CSF-1R^{MU} display eigenvalues twice as big as those of CSF-1R^{WT} (Fig. 22 A). In order to compare the eigenvectors from WT and MU, we performed an overlap, defined as a scalar product between the eigenvectors. Values showed a good agreement between CSF-1R^{WT} and CSF-1R^{MU} for modes 2 and 3 (Fig. 22 B).

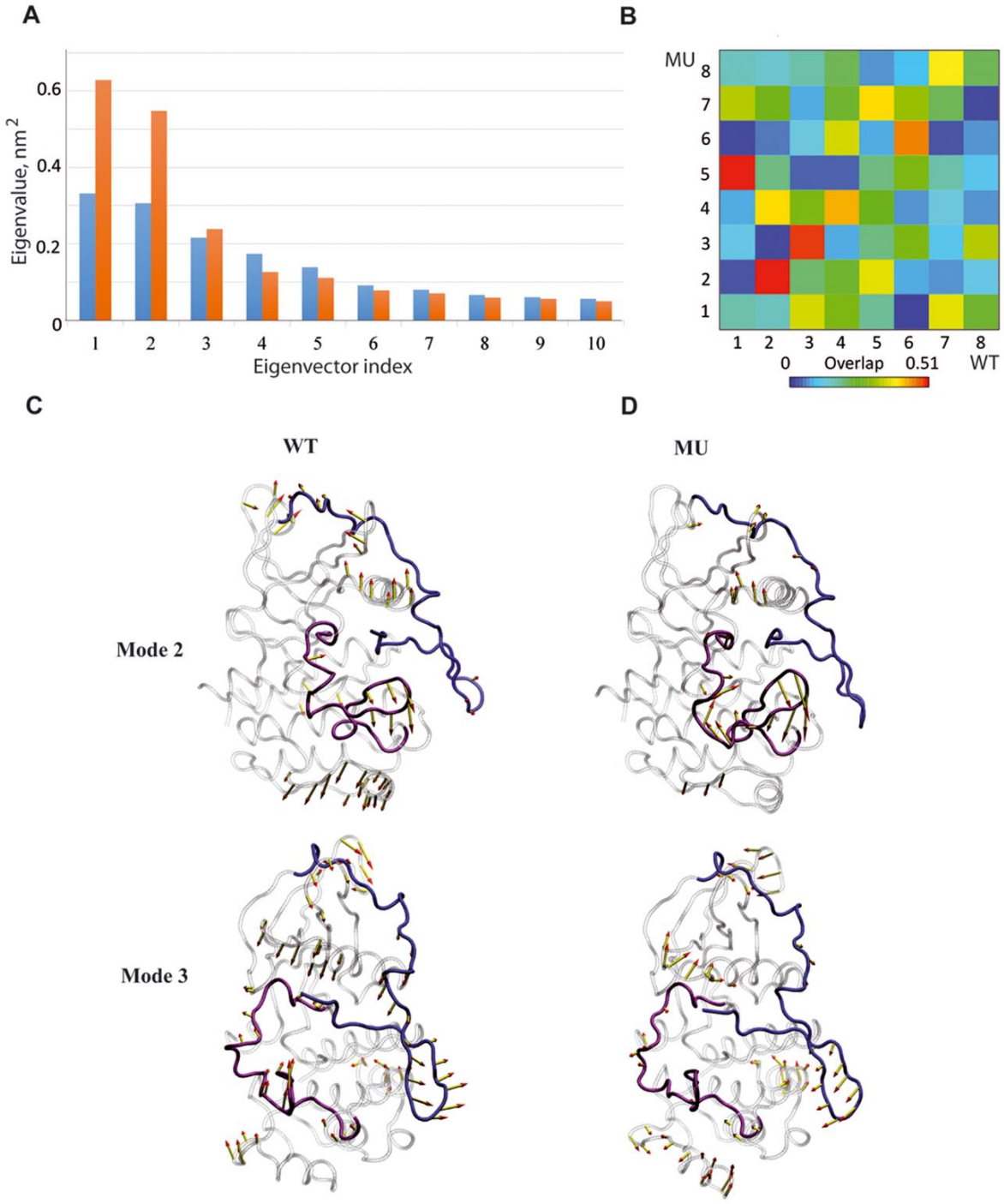


Figure 22: ***Principal component analysis (PCA) of CSF-1R cytoplasmic region in the inactive state.*** *The calculation was performed on the backbone atoms of CSF-1R^{WT}and CSF-1R^{MU}. Top: (**A**) The bar plot gives the eigenvalues spectra of CSF-1R^{WT}(blue) and CSF-1R^{MU} (orange) in descending order. (**B**) The grid gives the overlap between the first 10 eigenvectors from CSF-1R^{WT} (columns) and CSF-1R^{MU} (rows). The overlap between two eigenvectors is evaluated as their scalar product and represented by colored rectangles, from blue (0) through green and yellow to red (0.51). Bottom: Modes 2 and 3 atomic components for CSF-1R^{WT} (**C**) and CSF-1R^{MU} (**D**) are drawn as yellow arrows on the protein cartoon representation. JMR is in blue, A-loop is in violet and the rest of protein is in grey.* (DA SILVA FIGUEIREDO CELESTINO GOMES *et al.*, 2014)

In both CSF-1R^{WT} and CSF-1R^{MU}, the 2nd mode was associated mainly with the displacement of the A-loop, the loop linking β3-strand and Cα-helix and the C-terminus. Mode 3 showed the concerted movements of the loops connecting the β-sheet in the N-lobe and also movements in the proximity of the C-terminus, while we did not observed any movement in the TK domain that could be correlated to the JMR motions in both receptors. Noticeably the observed JMS motions in mode 3 depict "back-and-forward" movements in both models, which are not characteristic of JMR departure (Fig. 22 C-D), as it was observed for KIT^{D816V} mutant.

PCA has confirmed the previous analysis performed on the MD trajectories: no sign of JMR's departure from the TK domain, besides the absence of meaningful collective motions associated with the JMR. In order to obtain confirmation of this results, we have first used NMA, and after calculated the binding energy associated to the interaction between JMR and the TK domain.

Normal modes analysis (NMA) were performed on representative conformations of CSF-1R in the two forms. These representative conformations were issued from a convergence analysis applied on the concatenated MD replicas (LYMAN & ZUCKERMAN, 2006). The analysis is described in details at the Introduction (section 6.2.4) and Methodology (section 1.4) chapters.

Briefly, a set of *reference* structures were picked up randomly among the MD conformational ensemble of the trajectories and *reference* groups were composed of conformations from the two replicas of each trajectory. A good convergence quality was assessed when each *reference* structure was more or less equally represented in both replicas. To ensure the robustness of the method, we performed the analyses using five different random seeds for the *reference* structure picking up. For each form of the receptors, the fourth run containing the set of conformations that was better represented among the different replicas was chosen. The results of this analysis are summarized in Table 2 and Fig. 23.

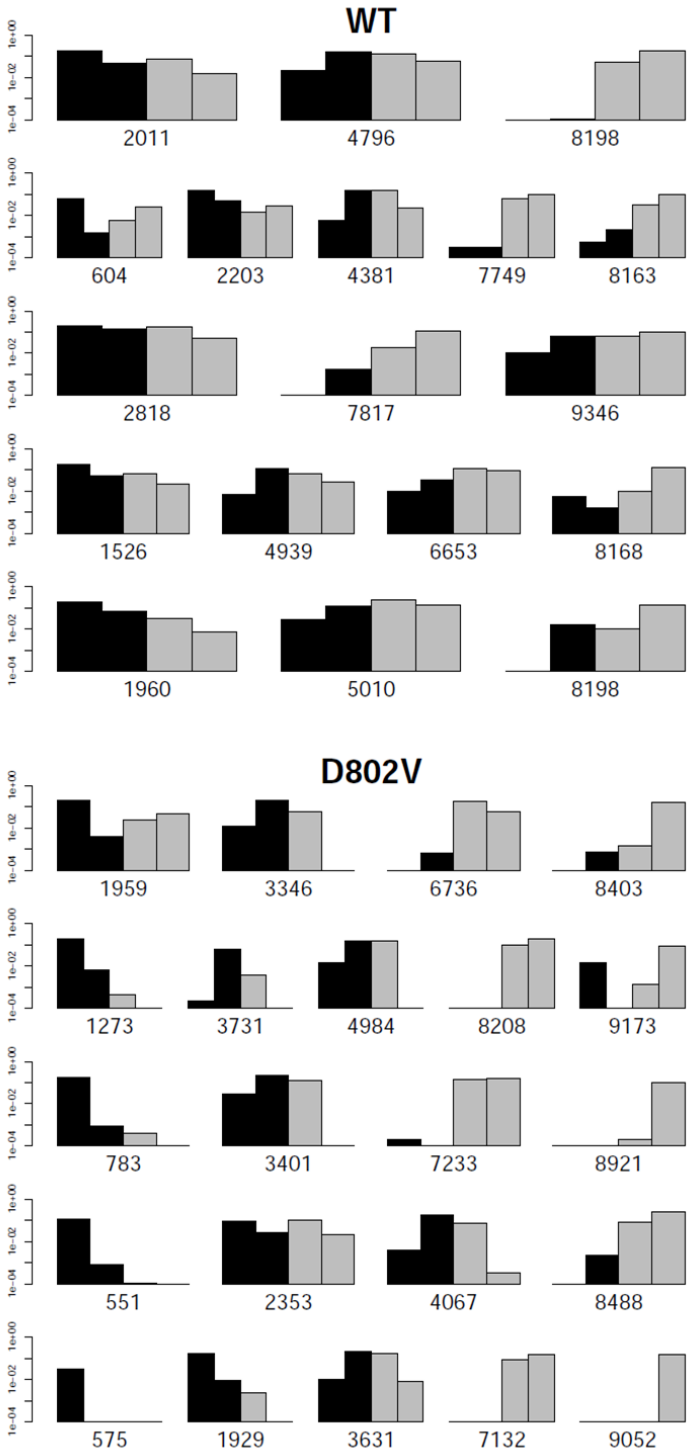


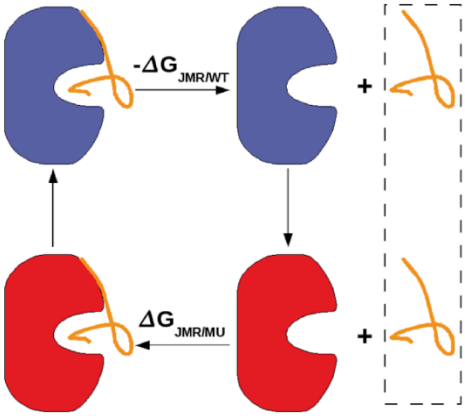
Figure 23: ***Convergence analysis of the MD simulations for CSF-1R^{WT} (WT) and CSF-1R^{MU} (D802V) models performed on the 90 ns concatenated trajectories.*** *Grouping of MD conformations was done using five independent runs calculated for each model. The populations of each group for each run are presented as histograms in the logarithmic scale denoted by different colors, black and grey from the 1st and 2nd halves of the two replicas, respectively. The identification numbers of each reference structure denotes the time (ns) in which it was picked from the MD trajectory. The fourth run of A and B contains reference structures that are better represented in both replicas and it was chosen for further NM calculations.*

Table 2: Parameters used in the convergence analysis of the native CSF-1R (WT) and its mutant (D802V) MD trajectories.
**best RMSD cutoff to represent the conformational diversity*
***range of reference structures identified in the 5 runs*
****number of runs with lone reference structures together with the range of lone reference structures in the five runs*

Parameter	**CSF-1R^{WT}**	**CSF-1R^{D802V}**
Cutoff r (in Å)*	1.8	1.8
Reference structures**	3-5	4-5
Lone reference structures***	3(1)	5(2-5)

When doing NMA, we were interested specifically on the degree of collectivity of the JMR atoms, k_{JMR}. The method of calculation is explained in the Methodology (section 1.6). The values of k_{JMR} range from *1/n* (only one atom among *n* is involved in the motion) to 1 (highly collective). The mean k_{JMR} value for CSF-1R^{WT} and CSF-1R^{MU} was of 0.44 and 0.42 respectively, indicating a low and statistically identical degree of collectivity in both proteins, denoting the absence of independent motions associated with the JMR. The visual inspection of the modes correlated to movements located at the JMR clearly indicated a great similarity between CSF-1R^{WT} and CSF-1R^{MU} (not shown). Altogether, the NMA analysis confirmed the absence of JMR displacement from the TK domain in the mutated protein, evidenced by the PCA.

The free energy of binding (ΔG) was computed on the individual conformations from MD simulations by the MMGBSA method. Although the large standard deviation values, results showed a tendency of JMR to display a lower affinity with the TK domain in CSF-1R^{MU} than in CSF-1R^{WT} (Fig. 24), similarly to previous observations in KIT^{D816V} (LAINE *et al.*, 2011). However, this difference was more pronounced in KIT ($\Delta\Delta G^{WT-MU}$= -42.68 kcal/mol)(LAINE *et al.*, 2011) than in CSF-1R, confirming the stronger coupling of JMR and the TK domain in CSF-1R. Such a coupling stabilizes the overall protein structure and dynamical behavior of CSF-1R evidenced by the low amplitude of the motions/fluctuations of JMR.



	ΔG CSF-1R^{WT}(kcal/mol)	ΔG CSF-1R ^{MU}(kcal/mol)
MD1	-62.87 +/- 14.91	-57.04 +/- 15.92
MD2	-62.21 +/- 13.89	-58.41 +/- 13.13

Figure 24: ***Binding energy changes between CSF-1R^{WT}and CSF-1R^{MU} in the inactive state.*** *Left : A thermodynamic cycle picturing the dissociation of JMR from KD in CSF-1R^{WT} (blue) and CSF-1R^{MU}(red). Right: The total free energy (ΔG) of the JMR binding to the kinase domain, computed over the individual MD simulations for both CSF-1R^{WT}and CSF-1R^{MU}.* (DA SILVA FIGUEIREDO CELESTINO GOMES *et al.*, 2014)

We also used the MD simulations data to calculate the occurrences of H-bonds involving key residues that maintain the inactive auto-inhibited form of CSF-1R (WALTER *et al.*, 2007). The H-bonds describing the contacts of the JMR and the A-loop residues with the residues from N- and C-lobes are summarized in Table 3 and illustrated in Fig. 25.

Surprisingly, the relative position of JM-B and TK domain residues in CSF-1R^{MU} appeared to be unfavorable to the H-bonds formation (Fig. 25 A). The occurrence of key H-bonds contributing to JMR anchoring to the TK domain, and to A-loop maintenance in an inactive conformation, were dramatically reduced in CSF-1R^{MU} (Tab. 3). The interaction between the JMR and the N-lobe, which is stabilized by an H-bond between Y546 (JM-B) and E633 (Cα-helix), was reduced by a factor of four in CSF-1R^{MU} compared to CSF-1R^{WT}. The occurrence of two other H-bonds, K545•••D625 and Y546•••E626, was reduced by a factor 2 in CSF-1R^{MU} compared to CSF-1R^{WT}. An alternative H-bond involving Y546 and D625 was detected in CSF-1R^{MU}, suggesting a partial compensatory effect.

Conversely, the H-bonds between the JMR and the catalytic loop from the C-lobe in CSF-1R^{MU} display none or only slight changes respectively to CSF-1R^{WT} (Tab. 3, Fig. 25 B). This observation indicates the strong coupling between the JMR and the C-lobe in both forms of receptor and correlates well with the highly conserved position of JMR respectively to kinase domain in CSF-1R^{WT} and CSF-1R^{MU}.

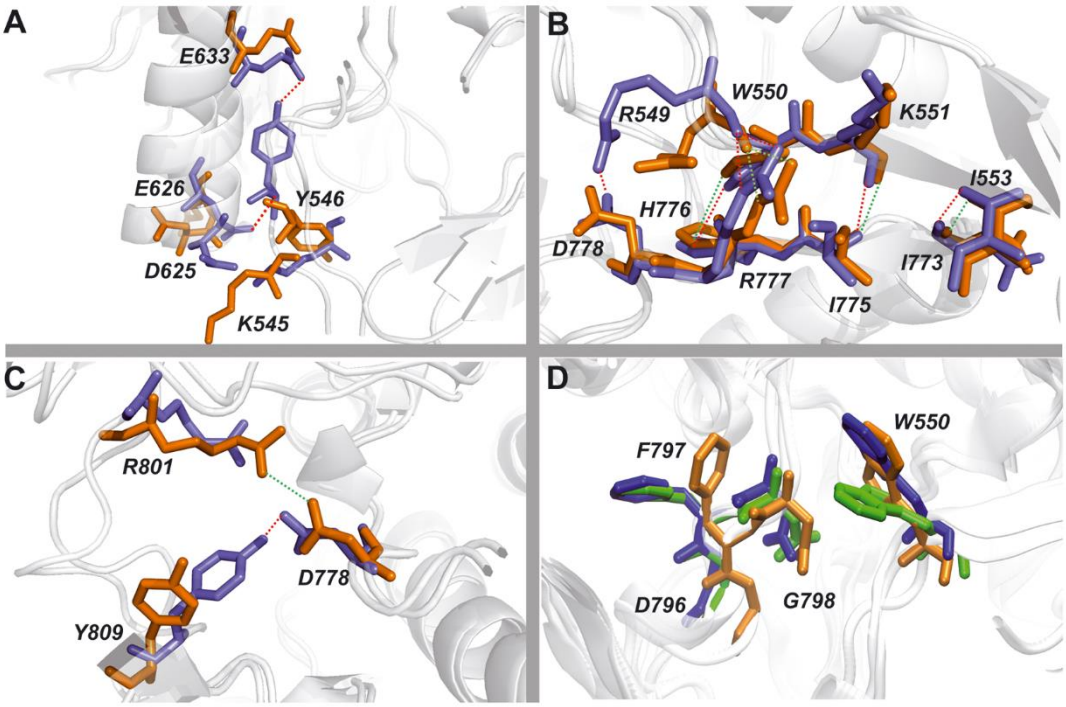


Figure 25: ***H-bond patterns in CSF-1R stabilizing the auto-inhibited inactive state of CSF-1R^{WT}and the non-inhibited inactive state of CSF-1R^{MU}.*** *H-bonds between residues from (A) JMR and Cα-helix; (B)JMR and C- loop and (C) A-loop and C-loop. Snapshots taken from the most representative conformations derived from MD simulations by the convergence analysis. All residues presented as sticks, in blue for CSF-1R^{WT}and in orange for CSF-1R^{MU}. The H-bonds are shown as dotted lines, red and green in CSF-1R^{WT}and CSF-1R^{MU} respectively. (D) The DFG motif conformation together with JMR's anchoring residue W550. Representation of DFG and W550 residues conformations originated from the crystallographic structure (2OGV, green) and representative MD conformations of CSF-1R^{WT} (blue) and CSF-1R^{MU} (orange).*

In addition to Y546, W550 is a crucial JM-B anchoring residue (WALTER *et al.*, 2007) that helps to hinder the active conformation of the A-loop by occupying the position that F797 (DFG motif) would acquire in the active form (MOL *et al.*, 2004). Representative structures derived from MD simulations showed a displacement of W550 side chain away from the ATP-binding site in CSF-1R^{WT} and CSF-1R^{MU}, when compared to its position in the crystallographic structure (Fig. 25 D). Remarkably, the DFG motif in CSF-1R^{MU} shows a conformational change in respect to CSF-1R^{WT} in the crystal and in the MD conformations (Fig. 25 D). All residues of the DFG motif in CSF-1R^{MU} are slightly displaced from their positions in CSF-1R^{WT}, and F797 side chain is pointed away from the ATP-binding site. Such position of F797 described as an "in" conformation the DFG motif that is specific for the inactive non-autoinhibited conformation of the receptor. The highly conserved residue F797, appears to serve as a conformational switch in the receptor.

Table 3: H-bonds stabilized the inactive conformation in CSF-1R^{WT} and CSF1R^{MU}. Residues involved in H-bonding and the H-bond occurrences (in %) are computed and averaged over MD simulations. Occurrences that showed a major difference are denoted by an asterisk.

JMR – C-helix contacts			**A-loop – C-lobe contacts**		
H-bond	CSF-1R^{WT}	CSF-1R^{MU}	H-bond	CSF-1R^{WT}	CSF-1R^{MU}
Y546•••E633	82*	19*	E825•••R900	100	100
K545•••D625	79	35	W821•••E847	100	100
Y546•••E626	68	38	W821•••S840	99	98
K545•••E628	-	30	Y809•••D778	82*	42*
T567•••K635	43	46	E825•••S636	79	68
K543•••E636	13	-	K820•••R855	63	61
Y546•••D625	-	30	R801•••R782	58	38
JMR– C-lobe contacts			D806•••R782	48	-
H-bond	CSF-1R^{WT}	CSF-1R^{MU}	K820•••N854	46	35
I553•••N773	100	100	Y809•••R782	44	33
R549•••R777	100	100	R801•••N783	34	26
K551•••I775	100	100	N808•••N854	20	29
R549•••D778	57*	-	R801•••N778	17*	82*
W550•••H776	54	38	P797•••N783	17	-
Y556•••V834	-	37			
Y556•••N773	21	21			
Y556•••Q835	-	20			

The A-loop inactive conformation was also stabilized by interaction of Y809 (A-loop) as a pseudo-substrate with the catalytic loop residue D778 (C-lobe) in CSF-1R^{WT} through the H-bond Y809•••D778, which is decreased by a factor of 2 in CSF-1R^{MU}. This destabilizing effect in mutant is compensated by H-bond R801•••D778, favored by the displacement of the R801 towards D778 (Fig. 25 C, Tab. 3). Despite this compensation, we lose a direct connection between the A-loop and the JMR through the H-bond network Y809•••D778•••R549, since we have the total loss of D778•••R549 in the mutant (Tab. 3).

Further, we compared the electrostatic potential surfaces of JMR and kinase domain in both receptors. The calculations were performed by the Adaptive Poisson-Boltzmann Solver (APBS) software using the crystallographic structures describing the inactive auto-inhibited state of the native receptors, CSF-1R (PDB ID: 2OGV, (WALTER *et al.*, 2007) and KIT 1T45, (MOL *et al.*, 2004)) receptors.

Although their structure is very similar, the two receptors display important sequence divergence in JMR. The JMR sequence contains 50 residues in KIT and 43 residues in CSF1R, including four basic residues in CSF-1R's JM-B & JM-S regions versus two in KIT; in addition, the distribution of nonpolar and polar residues is changed in these regions (Fig. 26 A). These subtle differences alter significantly the electrostatic surface of both proteins.

Particularly in CSF-1R, the charge complementarity between the JMR and the TK domain surfaces favors direct contacts between them (Fig. 26 B). Such profile in KIT shows relatively limited complementarity between the JMR and TK domain surfaces. The strong Coulomb interactions and the relevant H-bonds occurrences between JMR and kinase domain in CSF-1R confirm the tight coupling of these functional domains.

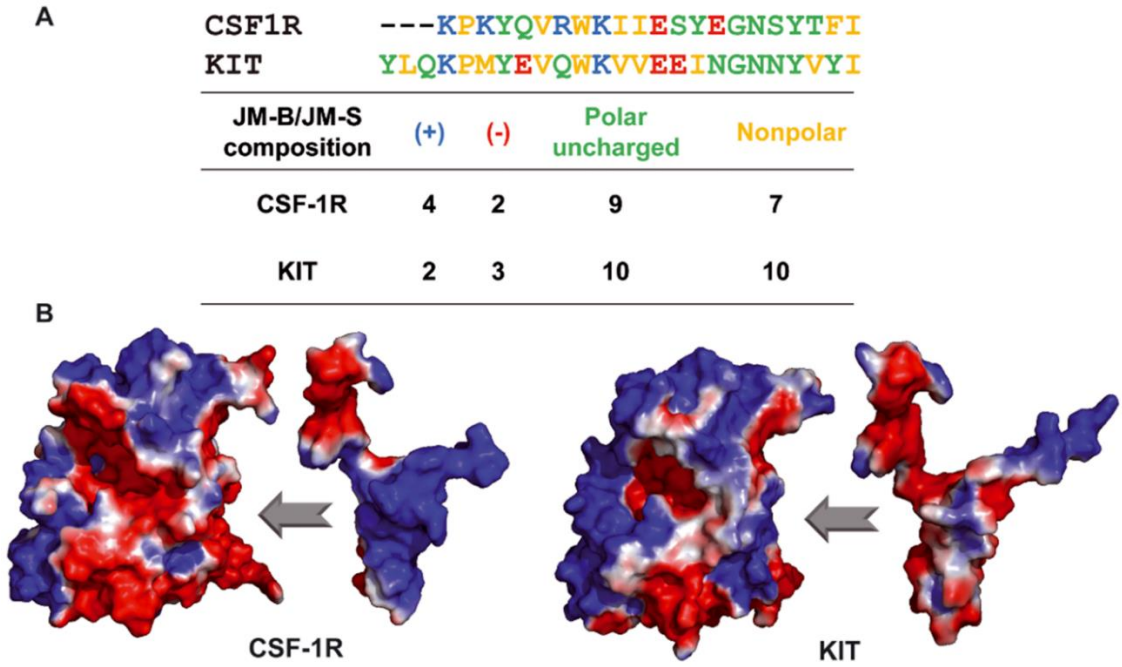


Figure 26: ***Comparison of the JMR sequence in CSF-1R and KIT and Electrostatic Potential (EP) surface in the two receptors.*** *(**A**) The amino acids composition of JMR (JM-B and JM-S) in CSF-1R and KIT. (**B**) The EP surface of TK domain and JMR in two receptors, CSF-1R and KIT. EP calculations on the Connolly solvent-accessible surfaces of the receptors were performed with the APBS software. The color scale ranges from red (electronegative potential) through white (neutral) to blue (electropositive potential).* (DA SILVA FIGUEIREDO CELESTINO GOMES *et al.*, 2014)

The JMR coupling with the TK domain controls the receptor activation process. Our group in France has recently developed a novel method, the *MOdular NETwork Analysis* (MONETA), designed for accurate characterization of communication pathways in a protein by exploring the inter-residues dynamical correlations computed from MD trajectories and the intramolecular non-bonded interactions (LAINE; AUCLAIR & TCHERTANOV, 2012).

Such approach applied to KIT put in evidence a well-established communication between the JMR and the A-loop tyrosine Y823 in KIT^{WT}, manifested as an extended network of H-bonds linking these two remote regions, through the catalytic loop D792 and Y823, linked in KIT^{WT} by a strong and stable H-bond, were identified as key residues in establishing of communication pathways. Destruction of this H-bond in KIT^{D816V} interrupted the allosteric coupling between these receptor segments leading to the structural changes in the JMR of KIT^{D816V} (LAINE; AUCLAIR & TCHERTANOV, 2012).

A study of CSF-1R using MONETA was performed to (i) analyze the communication pathways in the cytoplasmic domain of the receptor, (ii) evaluate the role of residue D802 in communication pathways and (iii) assess the impact of the D802V mutation on the protein internal communication network.

Identification of the protein regions representing the most striking local features of the two proteins' internal dynamics was carried out using a statistical technique known as Local Feature Analysis (LFA) (PENEV & ATICK, 1996) adapted from image processing to proteins (ZHANG & WRIGGERS, 2006). This method identifies clusters of residues named *Independent Dynamic Segments* (*IDSs*) that are formed around each *seed* and display concerted local atomic fluctuations and independent dynamical behavior (LAINE; AUCLAIR & TCHERTANOV, 2012).

The number of PCA modes retained for LFA was 17 in $CSF\text{-}1R^{WT}$ and 19 in $CSF\text{-}1R^{MU}$, the number of *IDSs* identified by MONETA being 8 in $CSF\text{-}1R^{WT}$ and 9 in $CSF\text{-}1R^{MU}$, respectively. The *IDSs* differences between the two receptors concern their feature, location, and size. To optimize the comparative analysis, the distinct *IDSs* were referred to as S_i, where i=1,2,…,N.

IDSs common to the two forms of receptor are located in JM-B (S1, residues 543-546), in JM-S (S2, residues 553-562 in $CSF\text{-}1R^{WT}$ and 554-562 in $CSF\text{-}1R^{MU}$), in the solvent-exposed loop that connects β2 and β3 (S3, residues 602-611) in the N-lobe, in the pseudo-KID (S4, residues

678-692), in the A-loop (S5, residues 810-817 in CSF-1R^{WT} and 809-817 in CSF-1R^{MU}), and in the C-terminal tail (S6, residues 914-922) (Fig. 27 A,B).

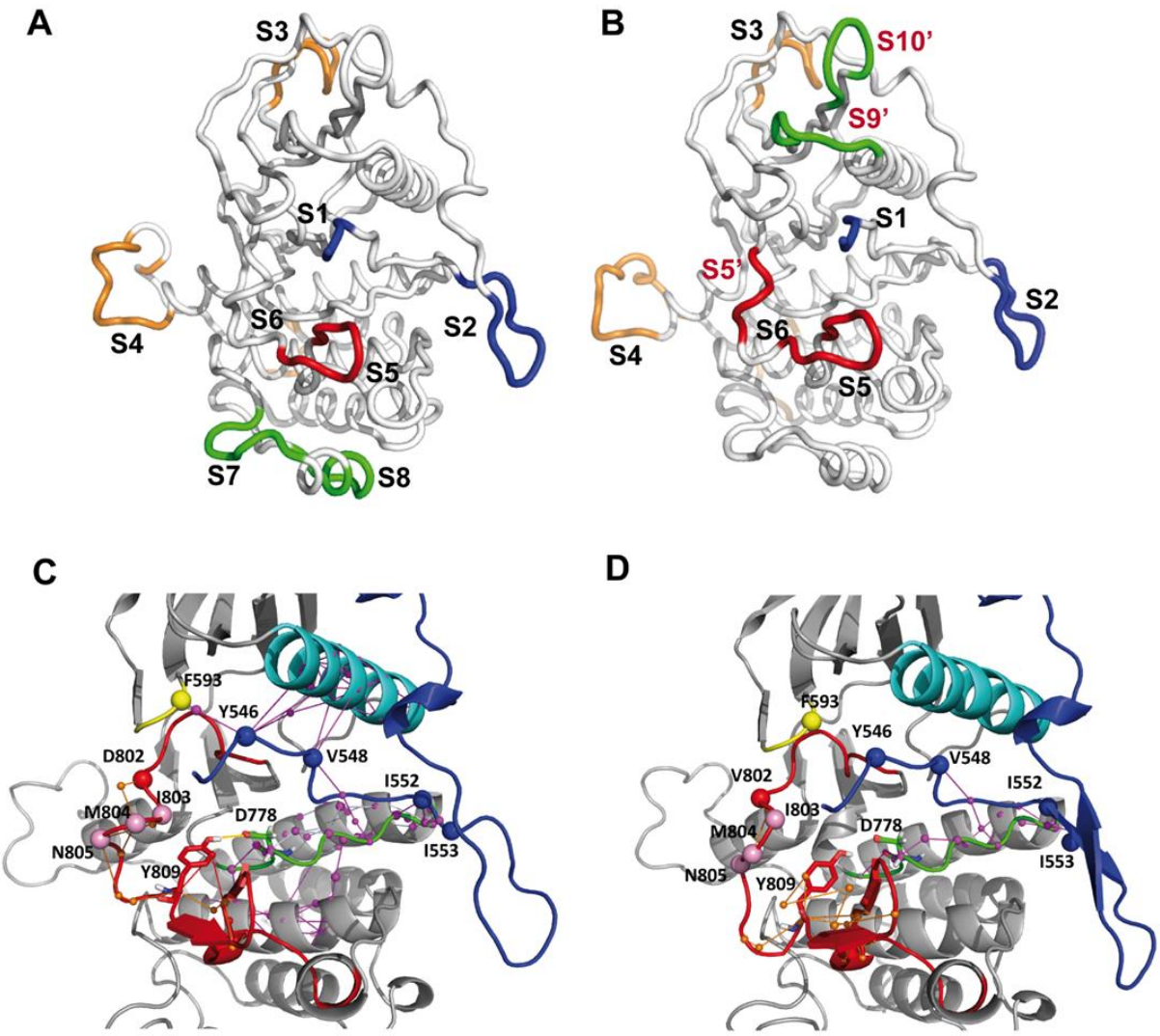


Figure 27: ***Independent dynamic segments and communication pathways in cytoplasmic region of CSF-1R.*** *Top: Structural mapping of the Independent Dynamic Segments (IDSs) identified in CSF-1R^{WT} (**A**) and CSF-1R^{MU} (**B**). The average conformations are presented as tubes. IDSs were identified from the analysis of the merged 60 ns concatenated trajectory. IDSs are referred to as S_i, where i = 1, 2,…,N, labeled and specified by color in the both proteins. The largely modified or newly found IDSs in the mutant are referred to as S'_i in red. Bottom: 3D structural mapping of the inter-residues communication in CSF-1R^{WT} (**C**) and CSF-1R^{MU} (**D**), computed over the last 30 ns of the individual MD simulations. MD 2 is taken for illustration. The average MD conformation is presented as cartoon. The proteins fragments are presented with different colors: JMR (blue), Cα-helix (cyan), P-loop (yellow), C-loop (green) and A-loop (red). Communication pathways (CPs) between residues atoms (small circles) are depicted by colored lines: CPs formed by the A-loop residues are represented in orange; by the JMR-residues in magenta. The key residues in the communication networks are labelled and depicted as bulky circles.* (DA SILVA FIGUEIREDO CELESTINO GOMES *et al.*, 2014).

The two *IDSs* specifically observed in CSF-1R^{WT} were found in the C-lobe (S7, residues 856-862 of the loop that connects Hα- and Gα-helices; S8, residues 867-874 in the G-helix). The three *IDSs* specifically observed in CSF-1R^{MU} were localized in the N-lobe (S9', residues 617-624 of the loop that connects β3 and Cα-helix; S10', residues 654-659 in the loop linking β4 and β5) and in the A-loop (S5', residues 802-806) (Fig. 27 A,B). Interestingly, the residues forming S9' in CSF-1R^{MU} were also found in S1, suggesting that the JM-B and the loop linking β3 and Cα-helix were associated in an entire self-reliant *IDS* (not shown). The other unexpected observations were the participation of D802V and Y809 in S5' and S5, respectively.

Using MONETA, we identified only one *IDS* in the N-lobe of CSF-1R^{WT} and three in that of KIT^{WT} (LAINE; AUCLAIR & TCHERTANOV, 2012), whereas *IDSs* in the JMR, the A-loop, the pseudo-KID, and the G-helix were identical in the two native receptors. The impact of the equivalent mutation on the *IDSs* in the cytoplasmic region of the two receptors is dissimilar.

In CSF-1R^{MU} three novel *IDSs*, S5', S9' and S10', are a consequence of increased concerted local motions of the A-loop and the loops linking β3 with Cα-helix, and β4 with β5 (Fig. 17). In KIT^{MU} such motion increase was observed only at the A-loop; the motions in two other loops were diminished respectively to KIT^{WT} (LAINE *et al.*, 2011). The two A-loop *IDSs*, S5 and S5', separated in CSF-1R^{MU}, were observed as superimposed and duplicated *IDSs* in KIT^{MU} (LAINE; AUCLAIR & TCHERTANOV, 2012). The two key residues, the point mutation and the A-loop tyrosine, are involved in *IDSs* (S5' and S5 respectively) in CSF-1R^{MU}, while in KIT^{MU}, only the point mutation is located in *IDS*.

Further, we studied the inter-residue communications linking different *IDSs*. To quantify the inter-residues communications, we computed the number of *communication pathways* (*CPs*) for each protein. In virtue of the strong influence of the dynamical behavior onto the communication pathways, the calculation of *CPs* was performed on the individual MD simulations. For instance, the communication network computed over the 60 ns concatenated trajectory contains 1692 and 1626 non-redundant paths in CSF-1R^{WT} and CSF-1R^{MU} respectively, indicating the mutation-induced diminishing of the communication network in the receptor (Tab. 4). Nevertheless, the total number of *CPs* can vary considerably among the different replicas for both forms (Tab. 4).

Table 4: Quantitative analysis of the communication network pattern among the different MD replicas. MD1, MD2 and MD12 are the two separate and merged trajectories respectively.
** Shortest paths = smallest paths involving two residues.*

Parameter	**CSF-1R^{WT}**			**CSF-1R^{MU}**		
	MD 1	*MD 2*	*MD 12*	*MD 1*	*MD 2*	*MD 12*
Shortest paths*	2082	2953	1692	2679	2341	1626
Hubs	39	66	30	57	48	36
Number of paths derived from A-loop residues						
D /V802	1	1		1	1	
Y809	3	5		5	9	
Number of shortest paths* connecting JMR to other functional segments						
JM-B– P-loop	1	1		0	0	
JM-B– Cα helix	0	17		3	1	
JMR– C-loop	24	27		39	21	

We were interested to investigate if the mutation D802V would compromise the communication between the receptor fragments determined as crucial in the activation mechanisms. Therefore, we looked for the *CPs* derived from the mutation site D(V)802, the A-loop tyrosine Y809 and the *CPs* that connect JMR residues to other functional TKD segments, such as the P-loop, the Cα-helix and the C-loop, all involved in the stabilization of the inactive auto-inhibited conformation of the JMR (Tab. 3).

Despite the variation of the number of paths and their communication profile among the two replicas for the same system, the data characterizing different forms of receptor indicates that the JMR communication, especially when involving the JM-B, is considerably affected in CSF-1R^{MU} respective to CSF-1R^{WT}. These data suggests that a local perturbation on the A-loop affects the JM-B communication with the P-loop and the Cα-helix, although JMR residues maintained a strong communication with the C-lobe, through the C-loop.

The differences in communication are illustrated using replica MD 2 for both CSF-1R^{WT} and CSF-1R^{MU}. The *communication pathways* identified by MONETA form either local small *CP* clusters or extended networks (Fig. 27C-D). In CSF-1R^{WT}, D802 is involved only in a local *CP*

protruded to M804 in the small 3_{10}-helix *H2* of A-loop, posterior to the mutation site. Y809 initiated short *CPs* with other A-loop residues, particularly with S807, L817, P818, V819 and W821. Similarly, to KIT^{WT}, no direct *CP* between the JMR and the A-loop in CSF-$1R^{WT}$ was identified. Nevertheless, the side chain of Y809 points toward the C-loop, probably as an effect of the H-bond Y809•••D778, highly prevalent during the MD simulations (Table 3).

Moreover, D778 in the C-loop is involved in a *CP* extended toward the JMR (Fig. 27 C). Consequently, this *CP* can transmit information from the JM-S residues forming *IDS* S2 to the catalytic (C-) loop residue D778, and further, through the H-bond Y809•••D778, to the A-loop residues. The JM-S residues are involved in distinct *CP* networks providing connection of the JMR to the other functionally crucial fragments of the kinase domain.

The well-established *communication pathways* formed by the JM-B residues (Y546 and V548) with the P-loop (F593) and the Cα-helix (residues 628-633), the extended *CPs* from the JMR residues reaching the C-loop, and the Eα- Fα- and Hα-helices, constitute a developed multi-branched *CP* network capable to coordinate the movements of N- and C-lobes involved in CSF-1R activation mechanisms, *i.e.* post-translational modifications and catalytic functions.

Interestingly, the *CPs* of each α-helix, Cα, Eα, Fα and Hα, are extended over the entire helix length, making a structurally preformed *communication fiber*. A considerable part of this extended *CP* network is completely lost in MD 2 from CSF-$1R^{MU}$, *i.e.,* no *CP* was observed between the JM-B and the P-loop, the Cα-, or the Hα-helices. Nevertheless, a relatively extended *CP* network is still observed between the JMR and the C-loop and the Eα- helix in CSF-$1R^{MU}$ (Fig. 27 D). This remaining network establishes communication between D778 and the JM-Switch but do not extend to the A-loop. Indeed, the H-bond Y809•••D778 controlling such CP extension in CSF-$1R^{WT}$, shows a two-fold diminished prevalence in CSF-$1R^{MU}$.

We also evidenced that, in CSF-1R, *communication pathways* connect S1 (JM-Binder) and S2 (JM-Switch) mainly to the molecular fragments not manifesting the concerted local atomic fluctuations (*IDSs*), except S5 formed by residues from the A-loop β-sheets. The links between residues belonging to *IDSs* and the other receptor fragments involved in *CPs* are held by H-bonds (Table 3). In CSF-$1R^{MU}$, the absence of H-bonds between the JM-B and the Cα-helix residues significantly altered *CP* profiles. Diminished occurrence of the H-bond Y809•••D778 provokes the *CP* interruption between V802 and Y809 which in CSF-$1R^{MU}$ are involved in S5

and S5' *IDSs,* respectively. By contrast, the conserved H-bond pattern between the JMR residues involved in S1 and S2 *IDSs* and the catalytic loop partially preserves the *CP* that links these *IDSs* with the C-lobe residues similarly to CSF-1R^{WT}.

Our analysis showed that despite a comparable pattern of *CPs* between the JMR and the A-loop in CSF-1R and KIT, their functional roles appear to be different. The established *CP* between the A-loop and the JMR through the catalytic (C-) loop is crucial for maintaining the allosteric regulation of the KD in KIT and its disruption in KIT^{MU} is a major contribution to its constitutive activation (LAINE; AUCLAIR & TCHERTANOV, 2012). Although, in CSF-1R, this disruption is only observed partially.

Another particularity of the CSF-1R communication pattern consists of the JMR communication with the glycine-rich P-loop and with the Cα-helix, not observed in KIT (Fig. 28). Mutual *CPs* of the JM-B residues with the Cα-helix are extended over the entire helix length in the native protein, while few and relatively small *CPs* are observed in KIT.

To search the origin of such difference in the two structurally similar receptors from the same RTK family having a considerable sequence identity, we pointed to the structural features of these receptors. Comparative inspection of the N-terminal domain structure in both receptors evidenced that position of the P-loop and the Cα-helix is (i) equivalent in the inactive state of both receptors; (ii) conserved over the inactive-to-active forms transition in CSF-1R; and (iii) highly dissimilar in KIT active and inactive forms (Fig. 29).

Indeed, the P-loop and the Cα-helix in the active state of KIT are shifted respectively to their positions in the inactive auto-inhibited state. The relative position of the P-loop and the Cα-helix in the active and inactive forms, which is equivalent in CSF-1R and divergent KIT, may reflect their different implication in the mechanisms regulating the activation of the two receptors. This hypothesis is coherent with the different communication pathways observed in the inactive auto-inhibited state of these receptors.

Nevertheless, it is important to mention that the crystallographic structure of CSF-1R active form (PDB ID: 3LCD, (MEYERS *et al.*, 2010)) was stabilized by a co-crystallized kinase inhibitor, while KIT active state structure (PDB ID:1PKG, (MOL *et al.*, 2003)) was reported with two phosphorylated tyrosine residues (Y568 and Y570) and with ADP bound in the active site.

These structural peculiarities suggest that displacement of the P-loop and the Cα-helix in KIT active state may be induced by phosphorylation events.

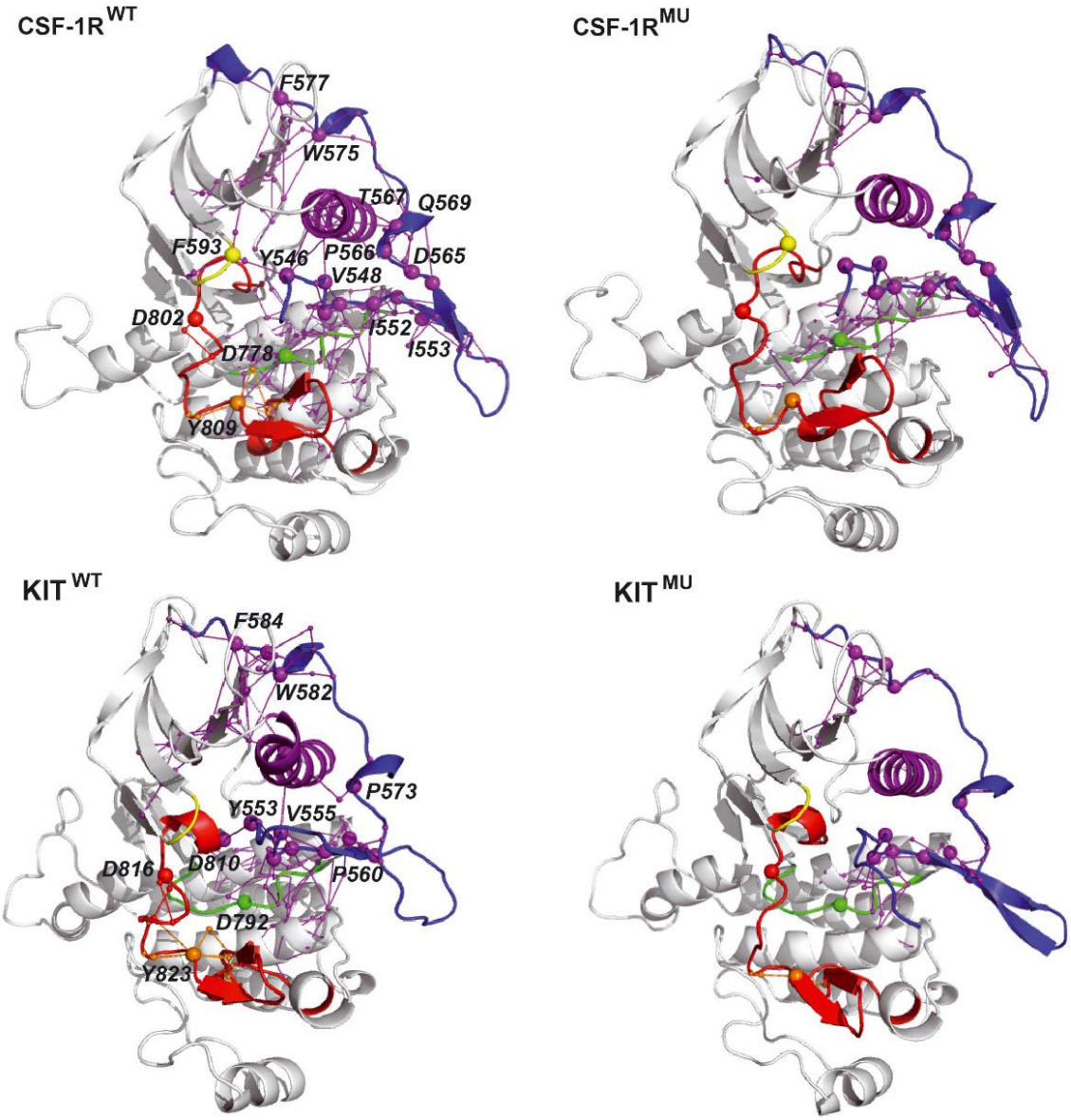


Figure 28: ***3D structural mapping of the inter-residues communication in CSF-1R^{WT}, CSF-1R^{MU}, KIT^{WT} and KIT^{MU}****. The average MD conformation is presented as cartoon. The proteins fragments are presented with different colors: JMR (blue), Cα-helix (violet), P-loop (yellow), C-loop (green) and A-loop (red). Communication pathways (CPs) between residues atoms (small circles) are depicted by coloured lines: CPs formed by the A-loop residues in orange; by the JMR-residues in magenta. The key residues in the communication networks are labelled (in WT) and depicted as bulky circles.* (DA SILVA FIGUEIREDO CELESTINO GOMES *et al.*, 2014)

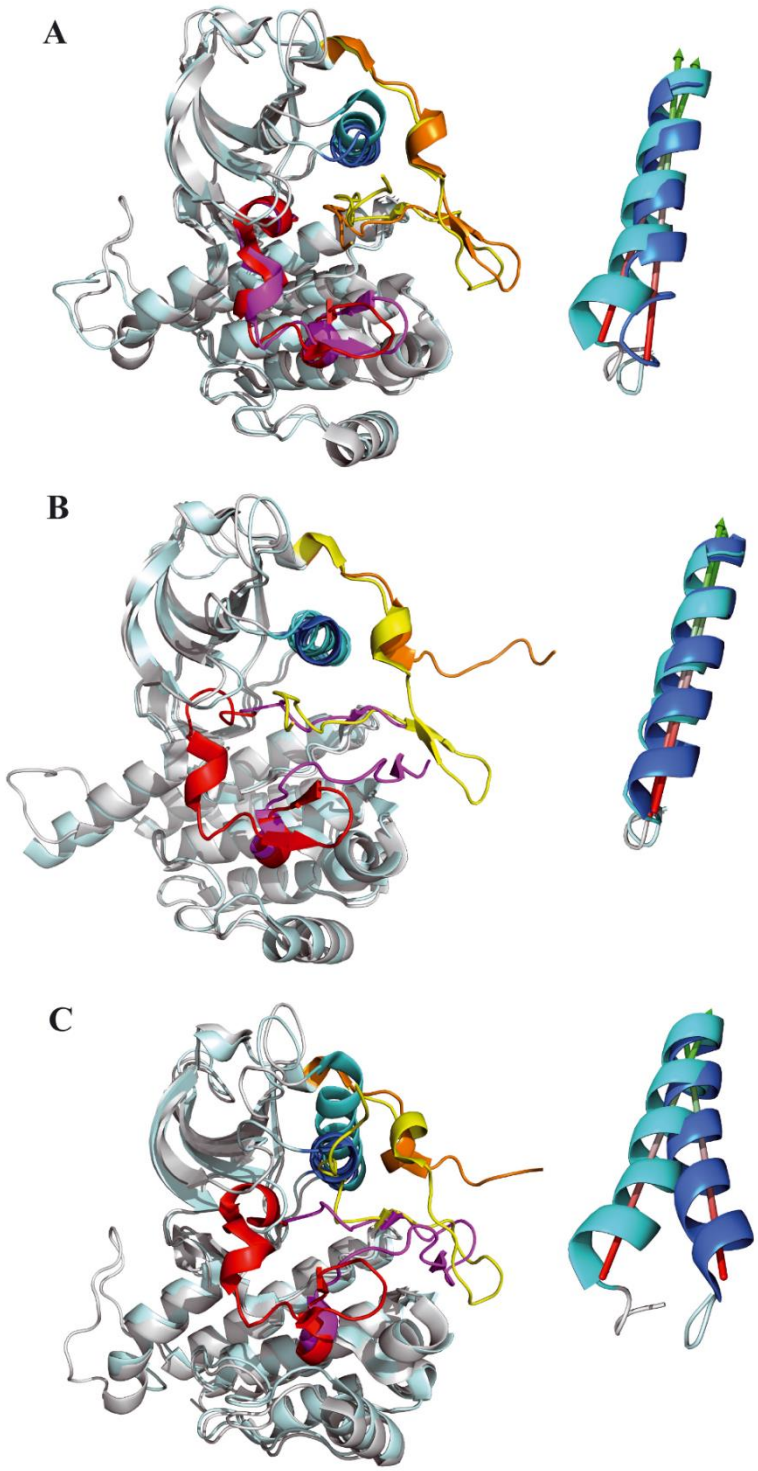


Figure 29: ***Structures of the cytoplasmic domain of CSF-1R and KIT in the native form****. Superimposition of the CSF-1R and KIT crystallographic structures: (**A**) CSF-1R (2OGV) and KIT (1T45) in the inactive conformation; (**B**) CSF-1R in the inactive (2OGV) and the active conformations (3LCD) ; (**C**) KIT in the inactive (1T45) and active (1PKG) conformations. The proteins are presented as cartoon, CSF-1R is in blue light and KIT is in grey light. The key structural fragments of receptors in the inactive and the active conformations are highlighted in color. The JMR is in yellow and in orange; the A-loop is in red and magenta; the Cα-helix is in cyan and blue. The relative orientation of the Cα-helix (inserts) in the two proteins is presented together with the principal axis of helices detected with PyMol.* (DA SILVA FIGUEIREDO CELESTINO GOMES *et al.*, 2014)

2. Imatinib binding mode to the CSF-1R and KIT receptors in their native and mutated forms

The second part of this thesis is dedicated to understand the resistance mechanism associated with the activating mutations on the receptors CSF-1R and KIT. The following mutations will be discussed here: D802V in CSF-1R; V560G, S628N and D816V in KIT. A 2D representation of imatinib's structure is shown in Figure 30.

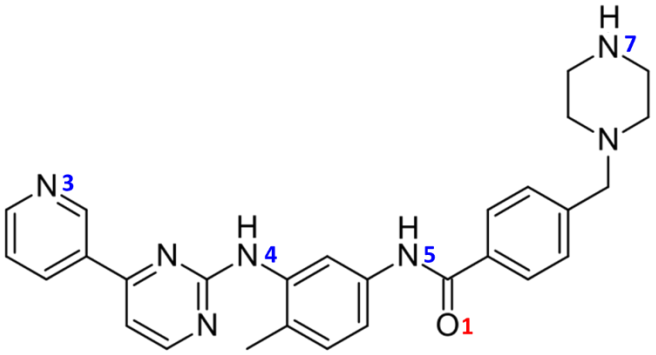


Figure 30: ***2D representation of the chemical structure of imatinib****. The labeled atoms in the figure constitute the ligand's atoms that are engaged in H-bonds interactions with the protein ATP-binding site residues. Hydrogen in N7 represents the protonation site.*

We decided to use equilibrated conformations for the WT and CSF-1R/KIT mutants derived from previous MD simulations (CHAUVOT DE BEAUCHÊNE *et al.*, 2014; DA SILVA FIGUEIREDO CELESTINO GOMES *et al.*, 2014). Protein conformations were selected based on a convergence analysis (see Methods). The purpose, though, was different from the use of this analysis in the part I of the Results.

The idea was to select *unique* conformations, one for each form of the receptors, to be used in the docking and MD simulations. By *unique*, we mean that they were visited more in each WT or mutant forms of KIT or CSF-1R, rather than been representative and sampled in all receptor's forms. We disposed of two MD replicas for each form of KIT or CSF-1R receptor, so we have concatenated all trajectories of KIT (WT and mutants together) and CSF-1R separately, based on the C-α atoms so we would not have topology problems associated with the different number of atoms in the mutants. Next, we applied the convergence analysis to the concatenated trajectory of KIT and the same was applied to the concatenated trajectory of CSF-1R. This time we discarded the representative conformations (spanned by all the

forms), but we were interested in the conformations more visited in the intervals corresponding to each WT or mutant's frames.

The next step consisted of docking imatinib into the chosen (one for each receptor form) conformations for WT and mutant forms of CSF-1R and KIT. We have not taken into account the JMR atoms in the moment of the convergence analysis, since they would be removed after the selection of the final conformations. The reason is explained in the next paragraph.

Although it is known that imatinib only binds the inactive form of Abl and KIT kinases (MOL *et al.*, 2004; NAGAR, 2007; SCHINDLER *et al.*, 2000; SEELIGER *et al.*, 2007), earlier attempts to dock imatinib into the auto-inhibited structure of KIT or CSF-1R have failed due to the buried conformation of the JMR. This is easily seen by the superposition of the inactive auto-inhibited structures of CSF-1R (PDB ID: 2OGV) and KIT (PDB ID: 1T45) with the inactive structure of KIT complexed with imatinib (PDB ID: 1T46). In the crystal 1T46, the JMR is not fully solved and it is displaced from its position in the auto-inhibited crystal 1T45, in which the residue W557 has its side chain oriented towards the ATP-binding site (Fig. 31). In the auto-inhibited structure of CSF-1R, we can observe the same behavior for the side chain of residue W550 (Fig. 31).

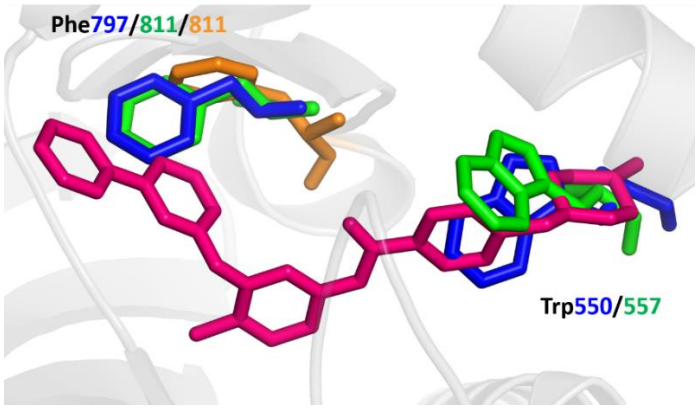


Figure 31: ***Imatinib bound to KIT in its inactive form (PDB ID: 1T46)****. Structures of auto-inhibited CSF-1R (PDB ID: 2OGV) and KIT (PDB ID: 1T45) were superimposed to highlight the side chain orientations of the residues Trp located at the JMR and the DFG-motif Phe. Imatinib is represented in sticks colored in pink, residues Trp and Phe from CSF-1R (550) and KIT (811) are colored in blue and green, respectively. In orange is represented the Phe from KIT when complexed with imatinib (1T46), the Trp is located on a missing part of the JMR in this crystal.*

Besides the W557, the F811 in the A-loop`s DFG motif is also inserted into the ATP-binding site in crystal 1T45 (Fig. 31). F811 would also impair the inhibitor binding due to its side chain position, which is displaced at the crystal 1T46. Same positioning of side chain is observed for F796 in crystal 2OGV from CSF-1R (Fig. 31). Therefore, to avoid the steric hindrances provoked

by the JMR domain, we extracted it from the selected conformations before performing the docking simulations and carefully looked the ATP-binding site surroundings to make sure that no steric hindrances were found.

Despite being unique, after the depletion (total or partial in case of KIT^{V560G}) of the JMR, all conformations are very similar concerning the ATP-binding site residues that interact directly with imatinib, described by (MOL *et al.*, 2004) (Fig. 32)

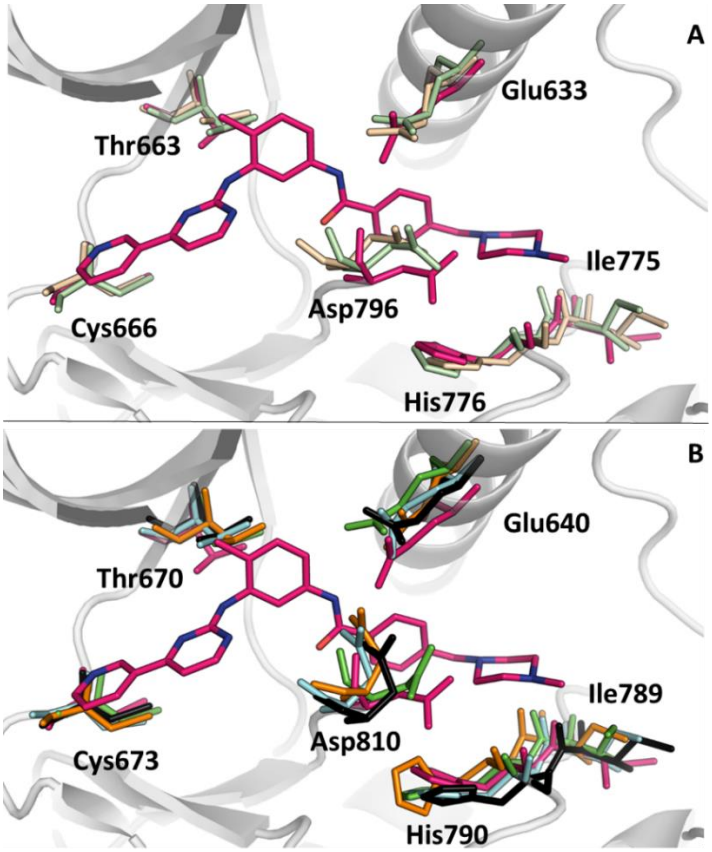


Figure 32: ***Selected conformations for the docking simulations****. Superimposition of the ATP-binding site residues, described at* (MOL *et al.*, 2004) *with their corresponding in the structures of CSF-1R (A) and KIT (B) selected by the convergence analysis and used in the docking simulations. Residues are represented in sticks and labeled following the numbering of CSF-1R and KIT separately. Imatinib and the ATP-binding site residues of crystal 1T46 are represented as sticks and colored in magenta. (A) Residues corresponding to* $CSF\text{-}1R^{WT}$ *and* $CSF\text{-}1R^{D802V}$ *are colored in pale green and wheat, respectively. (B) Residues corresponding to* KIT^{WT}, KIT^{V560G}, KIT^{S628N} *and* KIT^{D816V} *are colored in cyan, orange, green and black, respectively.*

Superimposition of the best docking poses for WT CSF-1R and KIT with the crystallographic structure of KIT complexed with imatinib (PDB ID: 1T46) (Fig. 33) shows a very good agreement of the ligand poses with imatinib's coordinates found in the crystal. This result reflects the validation for the docking experiments. In addition, the docking of imatinib

into KIT^{WT} reproduced the four H-bonds described in the crystal (PDB ID: 1T46) (MOL *et al.*, 2004). The docking poses for the CSF-1R and KIT mutants also showed a good positioning of imatinib in the ATP binding site of the receptors (Fig. 34).

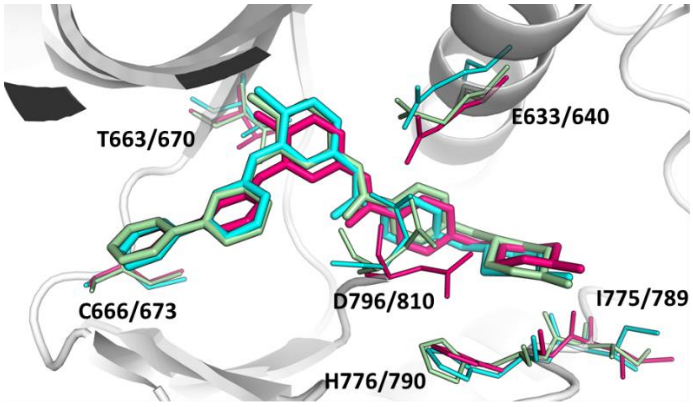


Figure 33: ***Validation of the docking methodology.*** *Superimposition of the docking best poses for WT CSF-1R (pale green) and KIT (cyan) with the crystallographic structure of WT KIT complexed with imatinib (magenta) referred by the PDB code 1T46* (MOL *et al.*, 2004)*. Imatinib and the surrounding ATP-binding residues are represented as sticks and the labels correspond to CSF-1R and KIT numbering, respectively. In the crystal, imatinib makes H-bonds with T670, C673, E640 and D810; H790 and I789 are described as hydrophobic contacts.*

Using the crystal as reference, the best poses presented RMSD values less than 1 Å (Table 5) and reproduction of four key H-bonds present at the crystallographic structure, with exception of H-bond between the imatinib and Thr663 in the CSF-1R^{WT} complex (Fig. 34). IFD scores are very similar among the different forms of the receptors. The extremely low energy IFD score for KIT^{V560G} is due to the constraints used during the docking simulation. Our first attempts of docking yielded bad results in terms of correct orientation of imatinib and energy, so, we have decided to apply a constraint that restricted the docking to a specified RMSD tolerance of 2.0 Å in respect to a reference structure, the crystallographic structure of KIT complexed with imatinib (PDB ID: 1T46).

The GlideScore values are coherent with the experimental data concerning the affinity with imatinib (Tab. 5), where KIT^{D816V} and CSF-1R^{D802V} are resistant (GAJIWALA *et al.*, 2009; TAYLOR *et al.*, 2006) and KIT^{V560G} is sensible, even more than KIT^{WT} (FROST *et al.*, 2002). The exception is the mutant KIT^{S628N} with values lower than those found for KIT^{WT} (Tab. 5). It is not very clear the role of this mutation in terms of sensitivity to imatinib (VITA *et al.*, 2014). According to the authors, the autophosphorylation of the mutant is abolished at a moderate

concentration (1µM), while for KIT^{D816V}, even with a concentration of 10 µM, we could still detect the autophosphorylation of the receptor (VITA *et al.*, 2014).

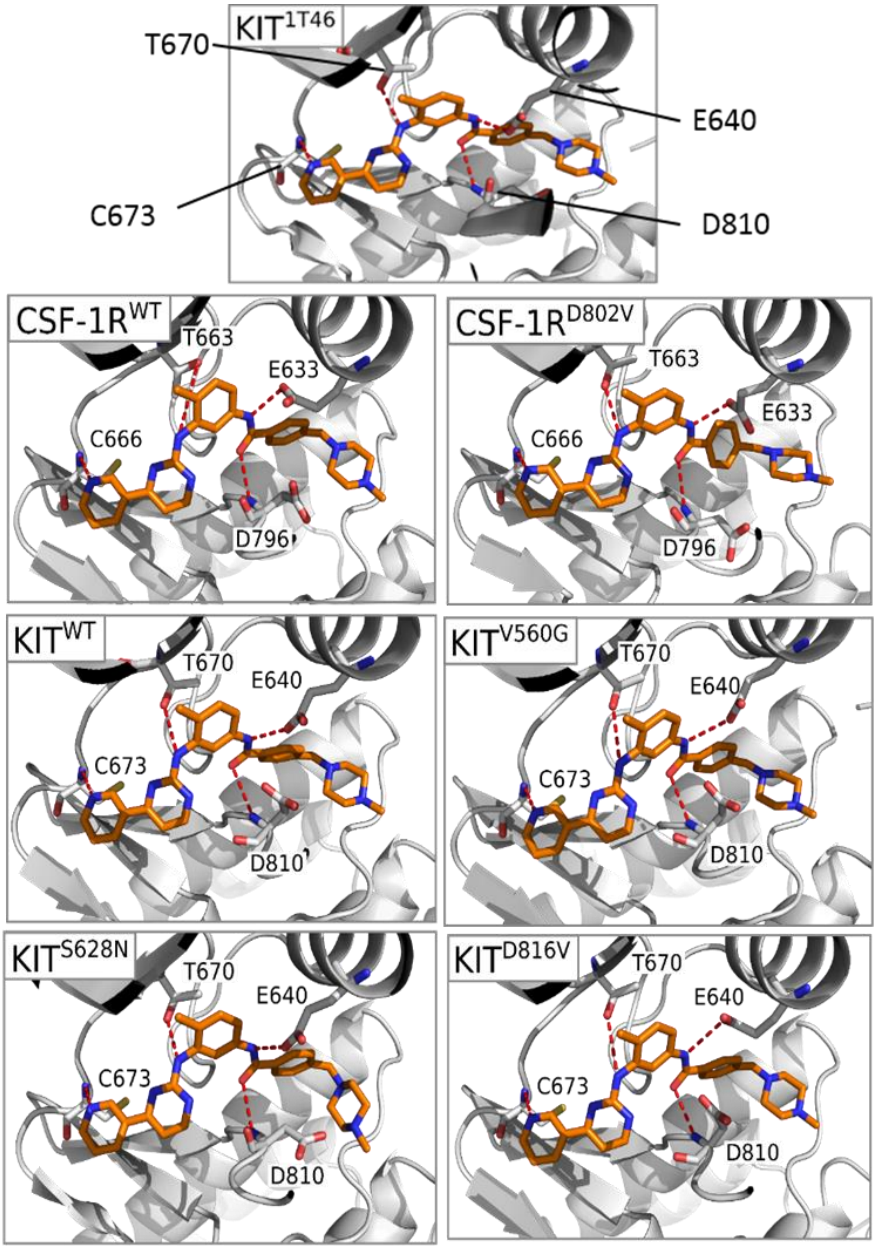


Figure 34: ***Best docking poses for each CSF-1R and KIT systems****. Imatinib is represented in orange sticks and the protein backbone is represented in grey as cartoon with the residues that interact with Imatinib in the crystal structure 1T46 are depicted represented as grey sticks and depicted in the first frame. Hydrogen bonds between the protein and the ligand are represented as dotted lines.*

Comparing the conformation of the ATP-binding site residues before and after the docking (Fig. 35), we probably could have used a rigid docking procedure, since the conformations are very similar, although, it is benefic to use the IFD to profit from a better accommodation of the ligand inside the binding site.

*Table 5: Score and RMSD of the best poses generated by the Induced Fit protocol. Maestro outputs the energy values in two scores: GlideScore and Induced Fit score (IFD). RMSD values were calculated using the crystal structure of inactive KIT complexed with imatinib (PDB ID: 1T46).*IFD score of KIT^{V560G} is extremely low due to the constriction used in the docking simulations (see Methods section).*

Protein	GlideScore (kcal/mol)	Score IFD (kcal/mol)	RMSD (Å)	IC_{50} (µM) of inhibition
KIT^{WT}	-7.8	-12.16	0.51	0.2
KIT^{V560G}	-9.2	-9857*	0.65	0.01
KIT^{S628N}	-8.6	-12.33	1.0	> 0.2
KIT^{D816V}	-7.0	-12.15	0.42	5
CSF-1R^{WT}	-7.9	-10.96	0.53	0.3
CSF-1R^{D802V}	-6.3	-12.39	0.52	> 4

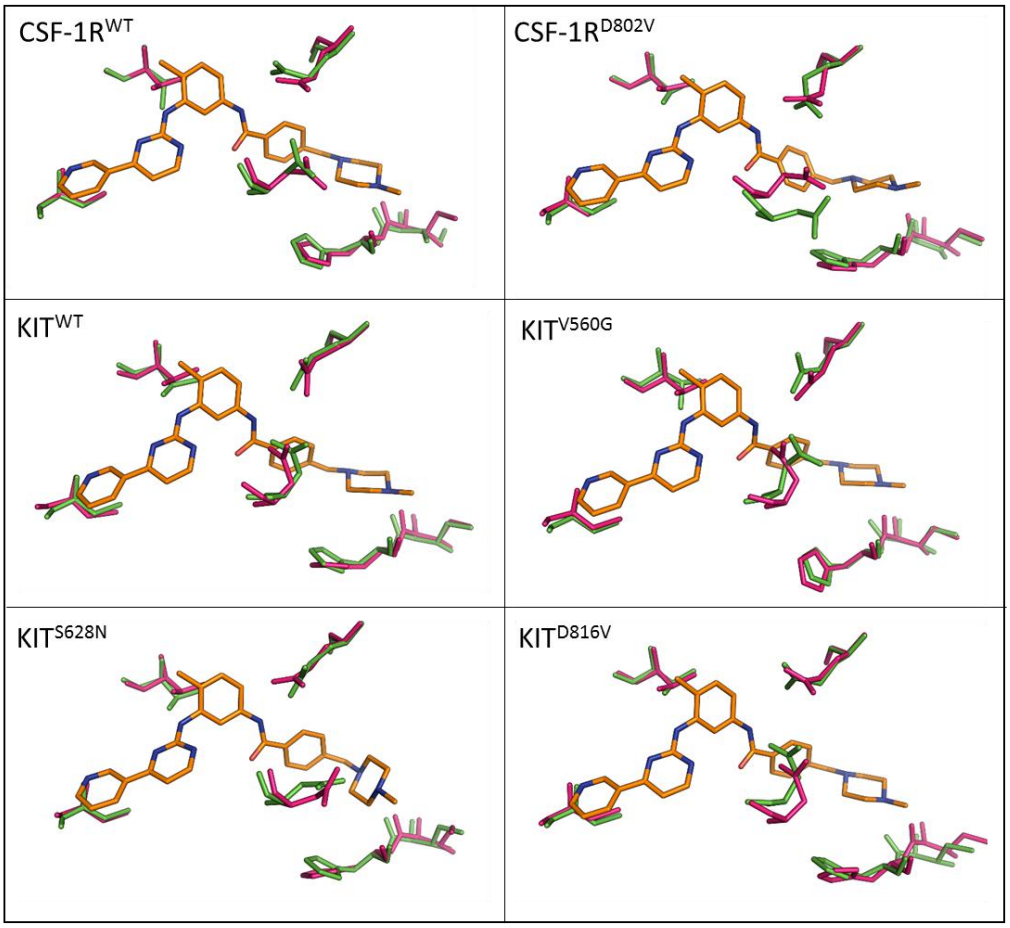


Figure 35: ***Comparison between the binding-site conformations pre- and after docking simulations****. ATP-binding site residues described in* (MOL *et al.*, 2004) *are represented as sticks and colored as magenta (before docking) and green (after docking). Imatinib is represented as orange sticks.*

Since we are interested in studying and understanding the differences in the affinity of different mutants with imatinib, the molecular docking runs were a first step of description. More accurate energy calculations are needed to better elucidate the problem. In addition, it would be valuable to comprehend the dynamic aspect of their interactions. That said, we selected the best pose of imatinib in each complex for further MD simulations. Further relaxation of the complexes is provided by MD simulations after molecular docking procedures, improving the protein-ligand fit.

MD simulations were performed in two replicas of 50 ns for each complex. The stability of the receptor-ligand complexes was measured by Root Mean Square Deviation (RMSD) calculation of the protein residues taking as reference the starting conformation of the MD simulations (Fig. 36). Two additional graphs were generated for the *apo* forms of KIT and CSF-1R in their WT state, e.g. in absence of imatinib (KIT^{apo} and $CSF\text{-}1R^{apo}$). The residues located at the C-terminal tail of CSF-1R and KIT are very flexible and responsible for significantly increasing the RMSD values, so they were not taking into account during this analysis.

RMSD profiles evidenced the great stability of the complexes, with RMSD values oscillating around 0.2 nm for most of them, with exception of KIT^{apo} and $CSF\text{-}1R^{apo}$, indicating that the presence of the ligand stabilizes the proteins. Exceptions are observed for one replica of $CSF\text{-}1R^{D802V}$, which presents increasing RMSD values after 30 ns of simulation, oscillating around 0.25nm and one replica of KIT^{D816V}, in which RMSD values reach 0.35 nm (Fig. 36). The change of the general net charge of the protein, caused by mutation, could have induced the conformational rearrangement.

The Root Mean Square Fluctuations (RMSF) shows the fluctuations over the backbone protein atoms in relation to an average conformation obtained from the two MD simulation replicas (Fig. 37). In order to facilitate the RMSF visualization, we have plotted the tridimensional representation of the RMSF in the form of b-factors (Fig. 38). Overall, both receptors in their different forms (WT and mutated) show a global protein stabilization.

The fluctuations are typically concentrated in residues located in loop regions of the proteins (Fig. 38). In particular, for KIT, the loop between the sheets β2 and β3 (**1**), the loop anteceding the αC-helix (**2**) (more pronounced for the mutant KIT^{S628N}, reaching some residues of the Cα-helix), the loop between sheets β4 and β5 (**3**), the A-loop (**4**,**5**), with different

extensions depending of the system and finally, the loop antecedent of the αG helix and part of αG helix itself (**6**).

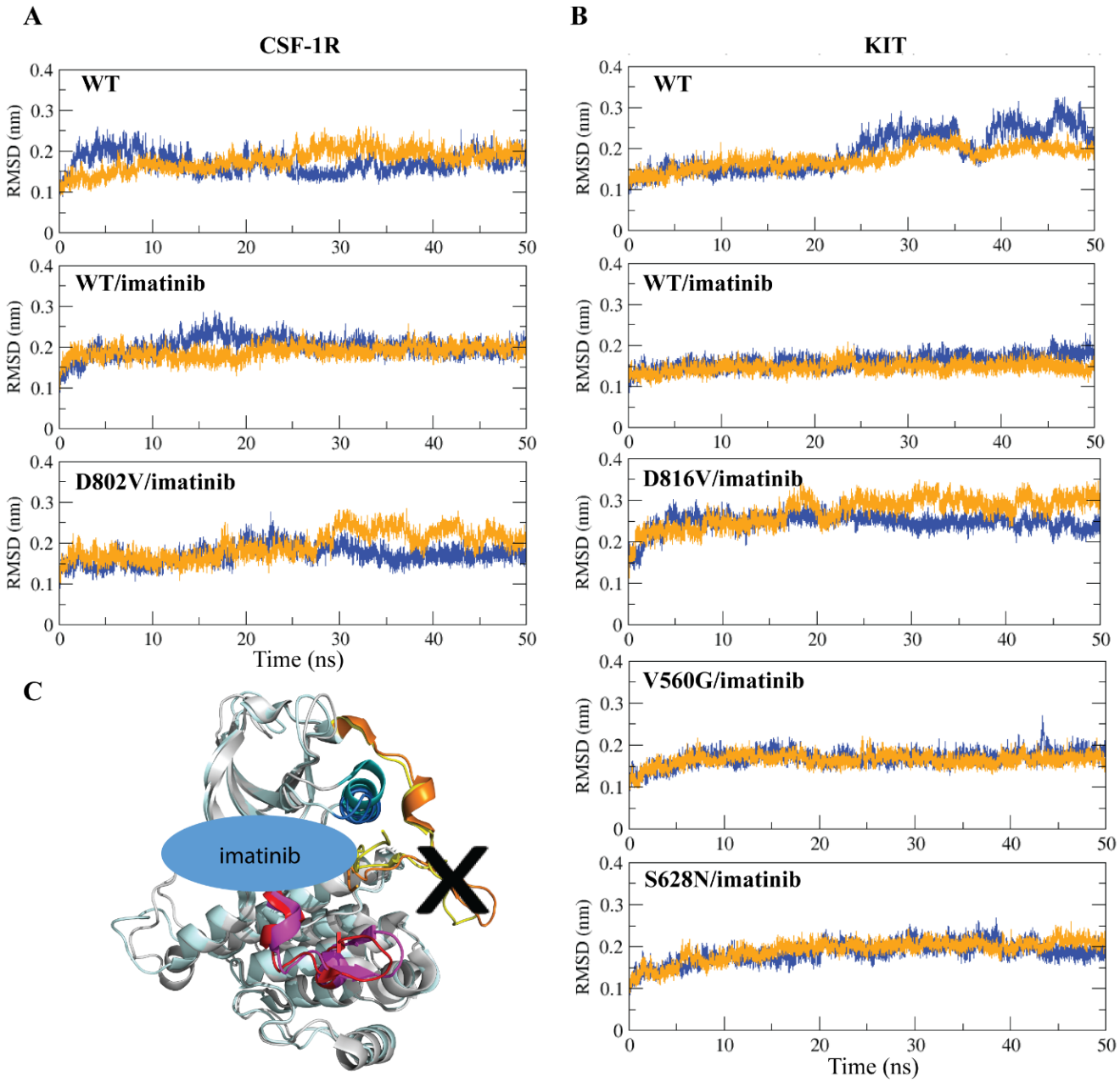


Figure 36: ***RMSD for the protein backbone atoms, excluding the C-ter tail.*** *RMSD values calculated for CSF-1R (A) and KIT (B) complexes in their WT apo, WT complexed with imatinib and mutant forms complexed with imatinib. The initial protein coordinates before the MD simulations were used as reference. Curves corresponding to replicas 1 and 2 are colored in blue and orange, respectively. (C) Superposition of the crystallographic structures of CSF-1R (2OGV) and KIT (1T45) as grey cartoon with key TK domain elements represented in different colors: JMR in orange and yellow, Cα-helix in blue and cyan, A-loop in pink and red, for CSF-1R and KIT, respectively. ATP-binding site where imatinib is placed is represented by an ellipse.*

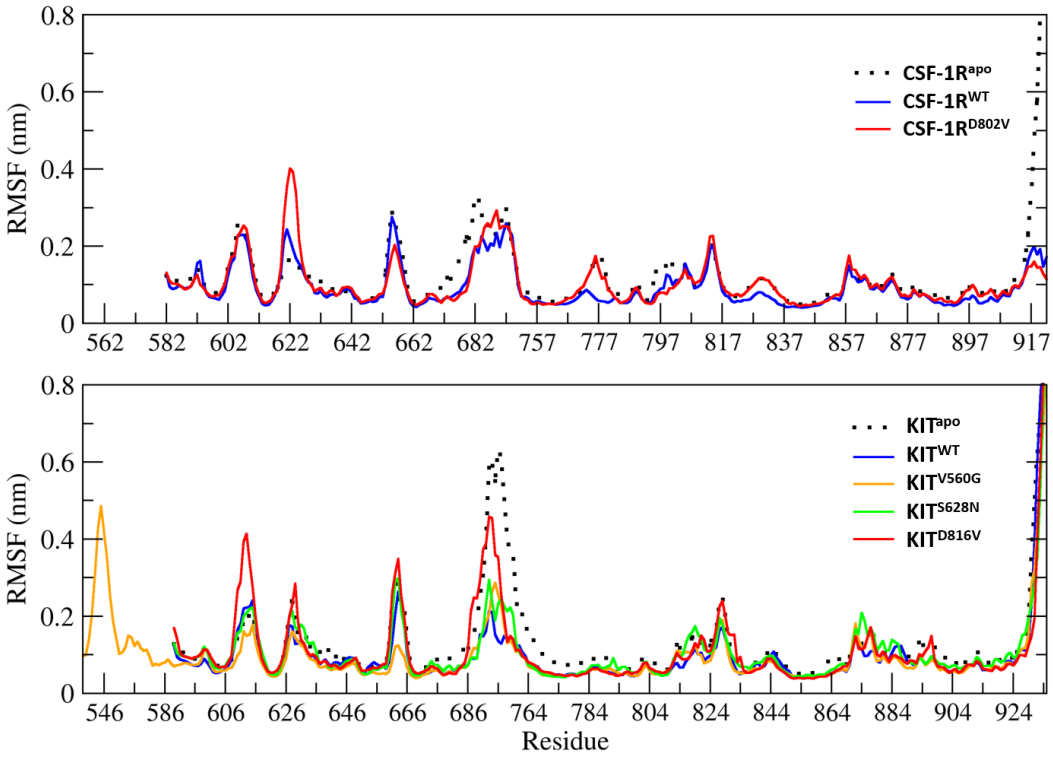


Figure 37: ***RMSF for protein backbone atoms****. RMSF values, averaged from both MD replicas, for CSF-1R (above) and KIT (below). The different forms of the receptor are designated in the legend.*

In CSF-1R, we find the same profile of fluctuation, with addition of P-loop residues in CSF-1R^{WT} (**7**) and, interestingly, the catalytic loop (C-loop) for CSF-1R^{D802V} (**8**). This enhanced fluctuation can be related to the H-bond pattern involving some of the C-loop residues.

In the WT form of CSF-1R, the A-loop residues Y809 and R801 make H-bond interactions with D778, which makes the A-loop "bind as pseudo-substrate" to the catalytic loop, contributing to maintain the inactive form of the receptor (WALTER *et al.*, 2007). H-bond profile calculated over the two MD replicas shows that in CSF-1R^{WT}, Asp778 is engaged in highly prevalent H-bonds with Asn783 (~100%), His776 (~60%), Arg801 (~78%) and Tyr809 (~39%), considering the simulation time corresponding to both MD replicas (Tab. 6).

These enhanced fluctuation on the mutant for this region of the catalytic loop could be a consequence of the change of the H-bond profile in CSF-1R^{D802V} for Asn783 (~70%) and for Arg801 (~1%), with a partial compensation in interactions with Tyr809 (~78%) and Arg782 (~94%) (Tab. 6).

Table 6: H-bond occurrences related to the P-loop residue D778 from CSF-1R. The values correspond to the average from both MD simulation replicas.

	H-bond occurence (%)	
	CSF-1R^{WT}	**CSF-1R^{D802V}**
D778···N783	100	70
D778···H776	60	67
D778···R801	78	1
D778···Y809	39	78
D778···R782	7	94

The H-bonds pattern between imatinib and ATP-binding site residues of WT KIT/CSF-1R confirm the stability of the four hydrogen bonds described at the crystallographic structure of KIT complexed with imatinib (MOL *et al.*, 2004)(PDB ID: 1T46) (Tab. 7) with addition of a newly observed contact with residue Ile789/775, which belongs to the catalytic loop. Together with the contacts observed between the hinge residue Cys673/666 and imatinib, these interactions are characterized by the high occurrence values (>95%), followed by the gatekeeper residue Thr670/663 with occurrence values varying between ~38-47%, DFG Asp810/796 with occurrence values varying between ~36-42% and the conserved Glu640/633 with occurrence values varying between ~12-30%.

KIT^{V560G} presents a very stable complex with imatinib, as expected since the mutation sensitizes the receptor to the inhibitor (FROST *et al.*, 2002). In addition, the H-bond occurrence values are slightly superior to the values found for KIT^{WT} for imatinib interactions with Asp810 (~73% against 42%) and Glu640 (~46% against 12%). For the resistant mutants, these values vary.

In the CSF-1R^{D802V} complex, a considerable decrease in the H-bond occurrence of imatinib with Ile775 (~18% against 96% for CSF-1R^{WT}) and a slightly reduction in the H-bond occurrence with Asp796 (~25% against 36% for CSF-1R^{WT}) are observed. Interactions of imatinib with Glu633 had similar values for both WT and mutant CSF-1R complexes.

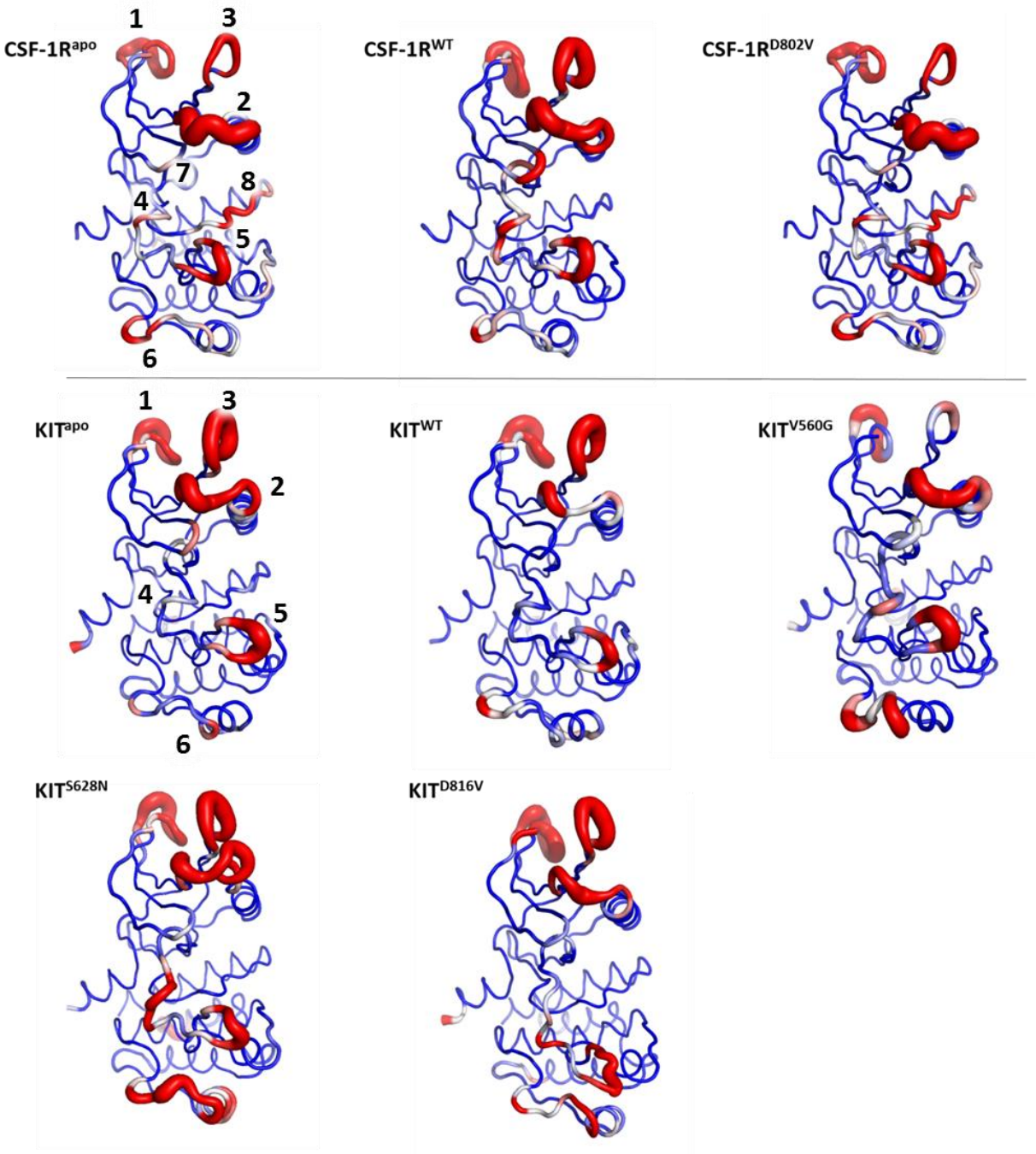


Figure 38: ***RMSF 3D representation on the protein backbone, excluding the KID and the C-ter regions****. The different conformations of CSF-1R and KIT are labeled. The regions that fluctuate the most are thicker, colored in red and numerated in the apo receptor forms.*

Table 7: Hydrogen bonds (H-bonds) occurrences between the targets and imatinib, averaged over the two MD replicas. The atom pairs for donor and acceptor interactions are depicted in the table. Imatinib atoms participating in the interaction are represented in figure 30.

	H-bond occurence (%)					
	N3-Cys666/N	**O1-Asp796/N**	**N7-Asp796/Oδ**	**N7-Ile775/O**	**N5-Glu633/Oε**	**N4-Thr663/Oγ**
CSF-1R^{WT}	98	37	0	96	34	38
CSF-1R^{D802V}	99	26	0	18	34	51
	N3-Cys673/N	**O1-Asp810/N**	**N7-Asp810/Oδ**	**N7-Ile789/O**	**N5-Glu640/Oε**	**N4-Thr670/Oγ**
KIT^{WT}	98	42	0	98	12	47
KIT^{V560G}	98	73	0	95	46	34
KIT^{S628N}	93	68	56	33	34	27
KIT^{D816V}	98	60	0	99	27	51

Surprisingly, the H-bond interaction pattern in the target-imatinib complexes formed by the resistant to imatinib KIT^{S628N} and KIT^{D816V} mutants, shows highly prevalent H-bonds, similarly to those stabilizing imatinib in the native KIT: inhibitor links to Cys673 (93/98%), Thr670 (27/51%), Glu640 (~34/27%) of KIT^{S628N}/KIT^{D816V} mutants. The main changes concerns residues Asp810 and Ile789. Imatinib in complex with KIT^{S628N} interacts differently, by alternating interactions with the backbone nitrogen of Asp810 and its two δ oxygen (Tab. 7).

Visual inspection of the trajectory showed that this occurs due to a flip of the methylpiperazin portion of imatinib. Also in consequence of this flip, the inhibitor loses interaction with Ile789, since this interaction involves the same imatinib's atom interacting with Asp810, the N7 nitrogen (Tab. 7). For KIT^{D816V} complex, the H-bond occurrences for Asp810 and Ile789 are highly prevalent (60/99 %).

The strength of the bio-molecular interactions involved in ligand binding/recognition by a target may be quantified by its binding free energy, and a range of computational approaches were developed to estimate this energies (KOLLMAN *et al.*, 2000; KOLLMAN, 1993; MACKERELL; BANAVALI & FOLOPPE, 2000; MEIROVITCH, 2007; MILLER *et al.*, 2012; PARENTI & RASTELLI, 2012; YTREBERG; SWENDSEN & ZUCKERMAN, 2006).

To distinguish between the factors which could contribute to the target's resistance or sensitivity associated with the studied mutations, we have further investigated and calculated

the binding's free energy of imatinib with the targets using the MM-PBSA approach (LEE; DUAN & KOLLMAN, 2000). The method combines three energetic components to account for the change in the free energy of binding (see Introduction, section 6.2.5). We have used a newly developed tool called *g_mmpbsa* (KUMARI *et al.*, 2014).

The analysis was performed for each imatinib-target complex over 9000 conformations derived from the concatenated trajectories, discarding from the calculation the first 5ns from each replica, also over the individual MD simulations and yielded similar results. Figure 39 shows the values for the calculation over the concatenated trajectories. The obtained results are coherent with the experimentally measured affinity of imatinib with the different receptors in their WT and mutant forms: KIT^{V560G}> KIT^{WT}/$CSF\text{-}1R^{WT}$ > KIT^{S628N} > KIT^{D816V}/$CSF\text{-}1R^{D802V}$ (Fig. 39).

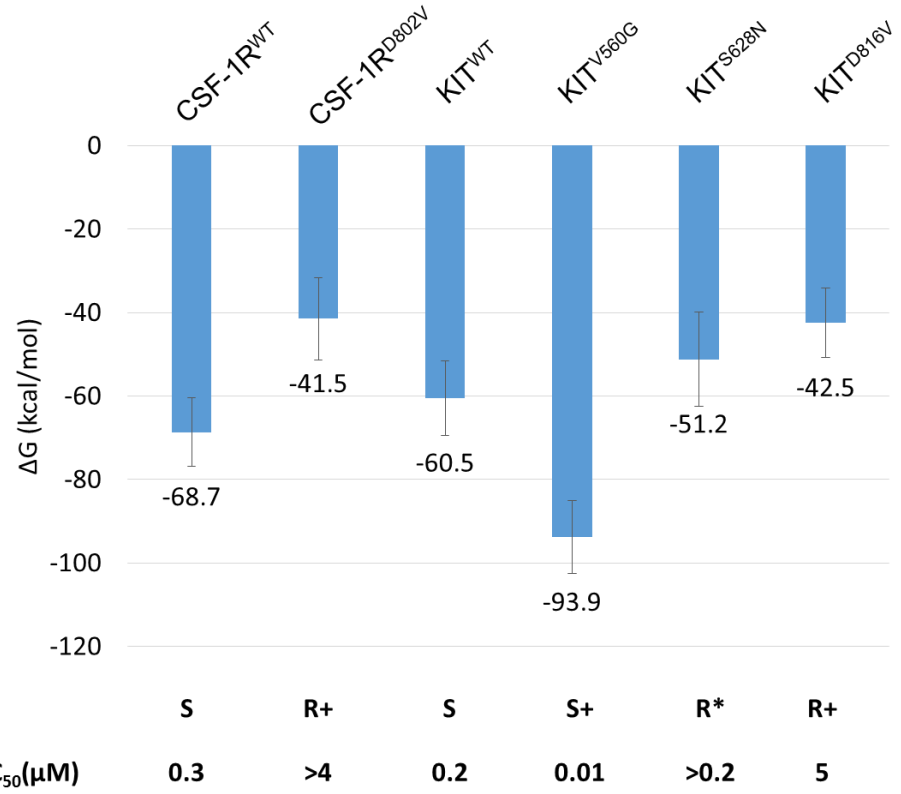


Figure 39: ***Binding energy between imatinib and the WT and mutant forms of CSF-1R and KIT****. Below the graph, the experimental IC_{50} values for the inhibition are indicated. The letters S and R indicate if the analyzed form of the receptor is sensitive or resistant to imatinib, respectively. The superscript + indicate more sensitivity or more resistance. The asterisk on the R associated with S628N mutation was placed because we do not know with certitude the resistance character of the mutation.*

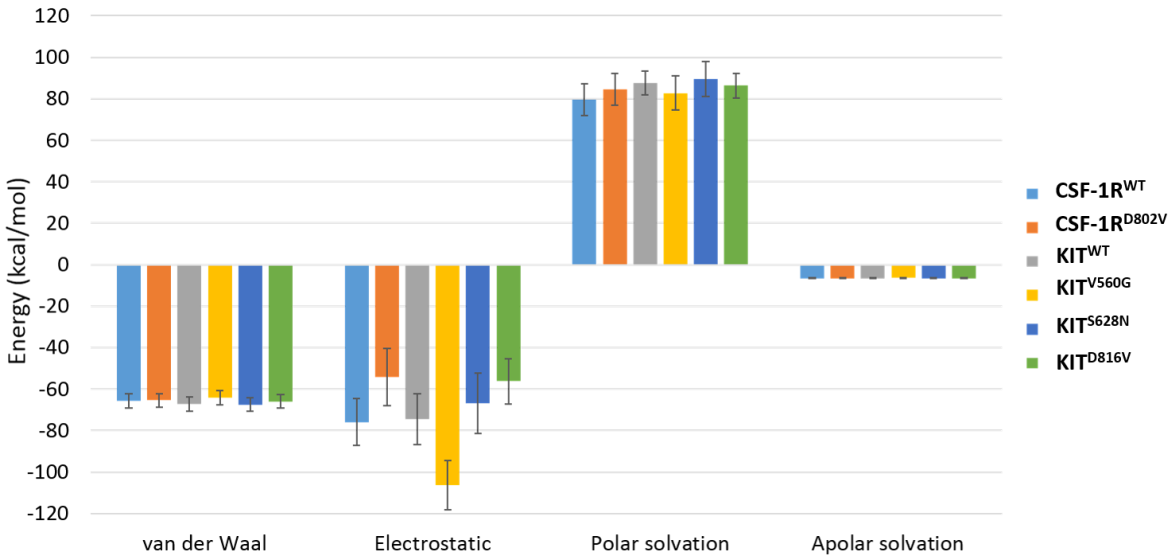


Figure 40: ***Binding energy decomposition****. The main components that contribute to the final binding energy, according to the MMPBSA approach, are represented.*

The experimental values shown in Fig. 39 and Table 4 were retrieved or approximated from the publications (in case of > or <). It is also important to mention that they were not retrieved from identical experiments protocols. For CSF-1R, the IC_{50} values were calculated from a test that measured the receptor phosphorylation as a reflex of the activation process. While in KIT, the IC_{50} values correspond to the inhibition of the proliferation of the studied cells that expressed the receptors. We are currently working on this problem by cooperating with an experimental biology laboratory, in which the researchers are going to repeat the same test, performed for CSF-1R, for KIT WT and mutant receptors. This will enrich our comparison.

The difference in affinity over the different mutants is intriguing since we did not observe such differences over the MD trajectories. By the energy decomposition into the different energy components, we noticed that the electrostatic energy term shows more variation among the other energy components (van der Waals, solvent polar and non-polar) and explains the difference in the final energy values between the WT and mutant complexes (Fig. 40).

Going further into the characterization of the energy components, we obtained the energy contribution for each protein residue (Tables 8 and 9). The negatively charged Asp and Glu

contribute the most, and the positively charged Lys and Arg residues are those that not only do not contribute but also are most unfavorable for the final binding energy. The favorable effect of negatively charged residues to the binding energy could be related with the protonation state of imatinib (charge +1).

Table 8: CSF-1R residues contribution to the final binding energy of imatinib. The residues belonging to the ATP-binding site are highlighted in yellow and the residue that can bear a mutation in orange. Not all residues are presented in this table, only the ones that contribute with absolute values superior to a cut-off of 4 kcal/mol. This rule does not apply for the ATP-binding site residues, since we are interested in their contribution. Std= standard deviation.

Residue	number	CSF-1R^{WT}	std	CSF-1R^{D802V}	std
LEU	582	5.72	0.17	5.68	0.22
LYS	586	6.26	0.37	6.29	0.32
LYS	595	6.91	0.26	6.95	0.36
GLU	598	-6.26	0.29	-6.25	0.29
LYS	606	4.48	0.27	4.47	0.29
GLU	607	-4.53	0.26	-4.53	0.28
ASP	608	-4.26	0.16	-4.21	0.16
LYS	612	5.31	0.20	5.30	0.19
LYS	616	15.95	2.47	18.29	2.01
LYS	619	7.54	0.74	7.55	0.88
ASP	625	-10.36	0.99	-9.44	1.40
GLU	626	-10.57	1.73	-9.23	1.45
LYS	627	7.39	0.32	7.61	0.51
GLU	628	-11.15	0.76	-11.31	1.09
GLU	633	-17.73	1.76	-18.73	2.10
LYS	635	11.60	0.88	12.06	1.11
THR	663	0.15	0.74	0.12	0.69
GLU	664	-5.81	0.55	-5.80	0.66
CYS	666	-2.26	0.48	-2.30	0.45
ASP	670	-8.59	0.30	-8.66	0.33
ARG	676	6.88	0.35	7.20	0.32
ARG	677	6.52	0.27	5.61	0.24
LYS	678	5.29	0.15	5.38	0.18
ARG	680	5.19	0.46	5.56	0.23
GLU	683	-5.87	0.37	-5.05	0.20
ASP	685	-4.53	0.19	-5.04	0.33
ARG	753	4.61	0.15	4.67	0.28
ASP	754	-5.32	0.12	-5.33	0.21
LYS	772	10.31	0.55	10.40	1.06
ILE	775	-4.24	0.75	-2.27	1.18
ARG	777	18.60	1.93	16.11	2.05

ASP	778	-14.80	1.28	-15.50	1.67
ARG	782	9.88	0.47	10.45	1.14
LYS	793	6.95	0.43	6.89	0.64
ASP	796	-18.08	1.57	-17.83	1.48
ARG	801	11.81	1.25	10.94	0.67
ASP/VAL	802	-8.66	0.68	-0.01	0.07
ASP	806	-8.68	0.53	-9.21	0.57
LYS	812	7.64	0.73	7.48	0.57
ARG	816	7.79	0.50	8.26	0.78
LYS	820	6.78	0.23	7.00	0.32
GLU	825	-8.78	0.20	-8.15	0.35
ASP	829	-7.82	0.22	-7.19	0.33
ASP	837	-15.15	0.61	-13.70	1.01
GLU	847	-8.11	0.25	-8.09	0.31
LYS	864	4.82	0.14	4.75	0.15
LYS	867	5.12	0.20	5.02	0.24
LYS	870	6.38	0.36	6.14	0.36
ASP	871	-5.40	0.17	-5.17	0.22
LYS	883	4.47	0.17	4.25	0.14
GLU	896	-6.45	0.31	-6.02	0.35
ARG	900	8.46	0.17	7.90	0.34
GLU	912	-5.14	0.14	-4.99	0.28
GLU	916	-4.36	0.14	-4.37	0.23
ASP	917	-4.22	0.12	-4.29	0.17
ARG	918	4.11	0.31	3.68	0.15
ARG	919	4.44	0.17	4.59	0.16
GLU	920	-4.42	0.27	-3.72	0.13
ARG	921	4.65	0.18	4.72	0.21
ASP	922	-8.83	0.29	-8.80	0.31

Table 9: KIT residues contribution to the final binding energy of imatinib. The residues belonging to the ATP-binding site are highlighted in yellow and the residues that can bear the mutation among the different systems, in orange. Not all residues are presented in this table, only the ones that contribute with absolute values superior to a cut-off of 4 kcal/mol. This rule does not apply for the ATP-binding site residues, or the mutation sites, since we are interested in their contribution. Residues containing an asterisk are present only at KIT^{V560G}. Std= standard deviation.

Residue	number	KIT^{WT}	std	KIT^{V560G}	std	KIT^{S628N}	std	KIT^{D816V}	std
GLY*	560			-0.33	0.07				
GLU*	561			-9.20	0.47				
GLU*	562			-8.25	0.60				
ASP*	572			-12.60	1.06				
ASP*	579			-8.69	0.49				

LYS*	581			7.94	0.42				
GLU*	583			-6.41	0.44				
ARG*	586			6.69	0.16				
ARG*	588			4.91	0.16				
LEU	589	5.65	0.18	0.05	0.02	5.88	0.23	5.67	0.17
LYS	593	6.21	0.26	6.17	0.28	6.37	0.34	6.20	0.26
LYS	602	6.92	0.21	6.94	0.21	7.09	0.25	6.90	0.23
GLU	605	-6.21	0.23	-6.21	0.24	-6.31	0.25	-6.22	0.21
ASP	615	-4.15	0.16	-4.21	0.15	-4.28	0.18	-4.21	0.21
LYS	623	15.22	1.64	17.36	1.69	15.71	1.84	17.54	1.79
LYS	626	7.37	0.38	7.72	0.49	7.67	0.43	7.27	0.56
SER/ASN	628	-0.07	0.06	-0.09	0.07	-0.07	0.10	-0.01	0.11
GLU	633	-10.78	0.93	-12.17	1.83	-10.69	0.99	-10.92	1.02
ARG	634	7.64	0.37	7.49	0.28	7.45	0.36	7.73	0.37
GLU	635	-11.14	0.90	-10.87	0.75	-10.97	1.13	-11.05	0.81
GLU	640	-16.69	1.65	-18.73	1.48	-18.19	1.95	-17.57	1.73
LYS	642	11.77	1.07	10.70	0.48	11.73	0.99	11.79	1.03
THR	670	0.15	0.70	0.41	0.70	0.48	0.79	0.18	0.69
GLU	671	-5.94	0.50	-6.08	0.47	-5.93	0.52	-6.07	0.44
CYS	673	-2.33	0.41	-2.22	0.41	-2.09	0.61	-2.24	0.40
ASP	677	-8.36	0.42	-8.28	0.30	-8.68	0.46	-8.58	0.28
ARG	683	6.88	0.30	6.76	0.29	7.09	0.29	6.90	0.35
ARG	684	5.74	0.16	5.75	0.19	5.94	0.28	6.02	0.28
LYS	685	5.18	0.10	5.19	0.11	5.26	0.14	5.17	0.11
ARG	686	5.35	0.20	5.43	0.24	5.56	0.27	5.47	0.32
ASP	687	-4.73	0.13	-4.76	0.12	-4.83	0.15	-4.88	0.27
LYS	693	3.57	0.15	4.25	0.23	3.57	0.16	3.75	0.17
GLU	758	-5.32	0.21	-5.28	0.27	-5.39	0.31	-4.84	0.38
ASP	759	-5.48	0.32	-5.38	0.33	-5.19	0.49	-5.39	0.28
ASP	760	-5.18	0.11	-5.19	0.11	-5.25	0.13	-5.18	0.10
GLU	761	-4.25	0.11	-4.23	0.10	-4.26	0.11	-4.23	0.11
ASP	765	-4.73	0.09	-4.77	0.09	-4.75	0.10	-4.77	0.08
GLU	767	-4.92	0.12	-4.99	0.13	-4.91	0.14	-4.96	0.13
ASP	768	-5.39	0.10	-5.44	0.10	-5.40	0.12	-5.43	0.10
LYS	778	6.83	0.26	7.01	0.29	6.71	0.25	6.90	0.30
LYS	786	10.06	0.42	9.92	0.39	9.93	0.41	10.16	0.43
ILE	789	-4.28	0.64	-4.40	0.77	-2.43	1.27	-4.31	0.74
ARG	791	19.31	1.84	18.63	1.60	21.19	1.50	18.12	1.63
ASP	792	-14.54	1.23	-16.48	1.26	-13.69	1.19	-14.23	0.83

ARG	796	9.55	0.37	9.18	0.41	9.91	0.53	9.83	0.33
ARG	804	5.24	0.10	5.29	0.10	5.29	0.14	5.30	0.10
LYS	807	6.84	0.41	7.12	0.37	7.01	0.41	7.10	0.36
ASP	810	-19.94	1.35	-18.25	1.28	-21.28	3.10	-18.78	1.31
ARG	815	13.47	0.83	10.43	0.71	11.97	1.16	11.56	1.10
ASP/VAL	816	-8.86	0.32	-8.58	0.42	-9.36	0.52	-0.17	0.10
LYS	818	8.22	0.56	8.14	0.52	8.51	0.72	7.18	0.51
ASP	820	-8.88	0.36	-8.58	0.37	-9.26	0.47	-9.25	0.40
LYS	826	6.99	0.34	7.26	0.57	6.98	0.35	7.28	0.70
ARG	830	8.15	0.45	7.68	0.44	8.31	0.43	7.89	0.57
LYS	834	6.92	0.20	6.83	0.20	7.07	0.25	7.17	0.30
GLU	839	-9.05	0.33	-9.14	0.22	-8.19	0.37	-9.18	0.23
GLU	849	-10.28	0.29	-10.24	0.25	-9.77	0.51	-10.25	0.26
ASP	851	-15.32	0.64	-15.35	0.48	-14.55	0.74	-15.53	0.56
GLU	861	-8.01	0.22	-8.01	0.21	-7.96	0.26	-8.08	0.19
ASP	876	-5.63	0.15	-6.03	0.27	-5.99	0.34	-6.19	0.25
LYS	881	5.00	0.14	4.94	0.13	5.02	0.22	4.94	0.13
LYS	884	6.20	0.33	6.11	0.28	6.09	0.36	6.11	0.33
GLU	885	-5.16	0.16	-5.01	0.11	-5.01	0.15	-5.03	0.12
ARG	888	6.28	0.27	6.48	0.11	6.07	0.28	6.49	0.12
GLU	893	-4.78	0.24	-4.76	0.20	-4.73	0.20	-4.75	0.21
GLU	898	-5.03	0.13	-5.12	0.17	-4.98	0.15	-5.09	0.13
ASP	901	-5.49	0.16	-5.54	0.15	-5.40	0.16	-5.45	0.15
LYS	904	5.21	0.13	5.26	0.12	5.14	0.12	5.22	0.12
ASP	908	-6.47	0.13	-6.52	0.11	-6.26	0.18	-6.50	0.12
ASP	910	-6.55	0.26	-6.54	0.22	-6.22	0.28	-6.58	0.24
LYS	913	6.02	0.15	6.05	0.13	5.80	0.19	6.03	0.14
ARG	914	8.56	0.22	8.61	0.15	8.01	0.30	8.63	0.17
LYS	918	7.96	0.64	8.22	0.61	7.77	0.61	7.90	0.62
GLU	925	-5.95	0.19	-5.98	0.19	-5.83	0.21	-5.86	0.17
LYS	926	5.12	0.21	5.22	0.21	5.01	0.23	5.04	0.14
GLU	930	-4.36	0.21	-4.50	0.42	-4.51	0.15	-4.60	0.13
ILE	935	-4.68	0.82	-5.19	0.90	-4.10	0.53	-3.99	0.32

Even the mutation point being outside of the ATP binding pocket, the substitution of an Asp for a Val in the mutants KIT^{D816V} and $CSF\text{-}1R^{D802V}$, has a significant consequence on the energy contribution: -8.68 kcal/mol for $CSF\text{-}1R^{WT}$ against -0.01 kcal/mol for $CSF\text{-}1R^{D802V}$; -8.58 kcal/mol for KIT^{WT} against -0.17 kcal/mol for KIT^{D816V} (Tables 8 and 9). In the case of KIT^{V560G},

the presence of Gly in the mutation site does not interfere much in terms of energy contribution (-0.33 kcal/mol). The same effect is observed for KIT^{S628N} (Tab. 9)

Analyzing the ATP-binding site residues, they have very similar energy contributions to imatinib binding energy in CSF-1R and KIT. Since the ATP-binding site residues show very small differences in energy concerning WT and mutant systems, the contribution of the protein residues at the vicinity of the ATP-binding site could interfere in the global binding energy. To facilitate the visualization of the residues that have the most significant changes in energy, we have subtracted the mutant contributions from the WT (Figs. 41 and 42).

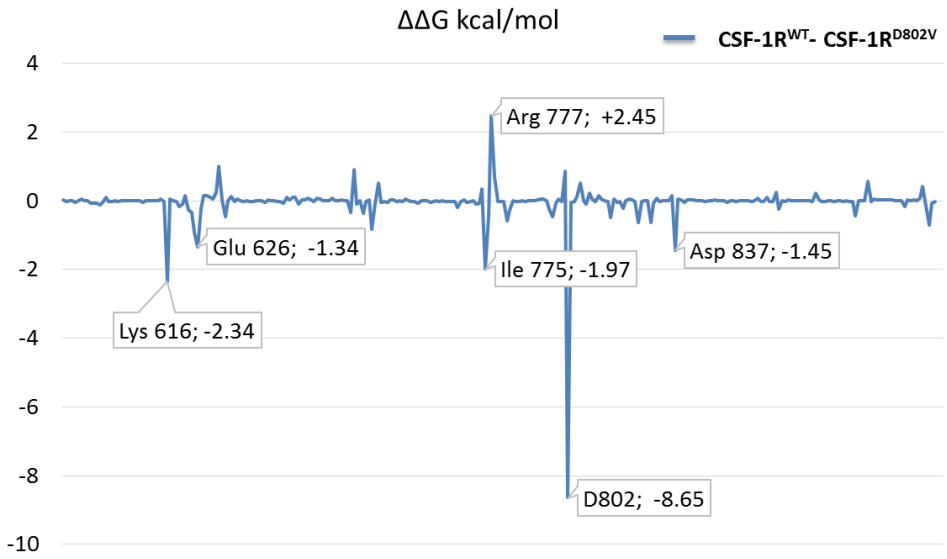


Figure 41: ***Difference between CSF-1R^{WT} and CSF-1R^{D802V} residue contribution to the binding energy****. Residues presenting differences superior to 1 or inferior to -1 kcal/mol are highlighted.*

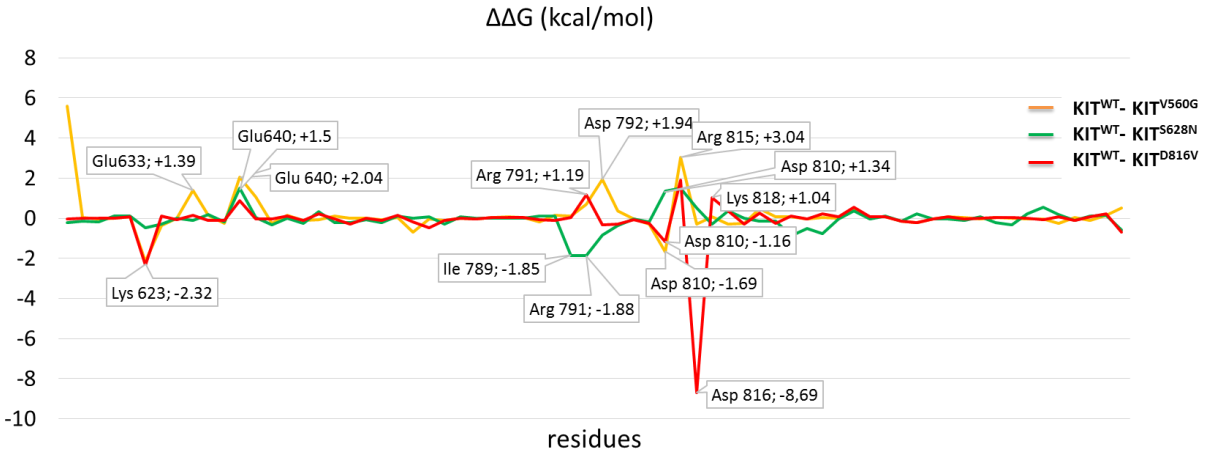


Figure 42: ***Difference between KIT^{WT} and each mutant for the residue contribution to the binding energy****. Residues presenting differences superior to 1 or inferior to -1 kcal/mol are highlighted. The energy contribution for the truncated portion of the JMR was not considered.*

Considering CSF-1R, besides Asp 802, Ile 775 is the only residue making direct contact with imatinib through H-bonds and the higher contribution in energy for the WT is consistent with the higher prevalence of this H-bond, in comparison with the mutant that has a reduction of ~ 80% in occurrence (Tab. 7). Other residues that show up with more differences are Lys 616 and Glu 626. They are not located at the ATP-binding site, but in the Cα-helix, close to the inhibitor.

Regarding KIT, when comparing WT and KIT^{V560G}, we see that the mutant has more residues contributing favorably to its binding energy, which could explain its higher affinity for imatinib. The complex formed by KIT^{S628N} shows more difference for two residues located at the C-loop: Ile 789 and Arg 791. During the MD simulations, the methylpiperazin portion of imatinib is flipping inside the ATP-binding site, this causes a loss of in H-bond interactions with Ile 789 (Tab. 7). As far as it concerns the complex formed by KIT^{D816V}, besides Asp816, the residues more unfavorable are Lys 623 and the ATP-binding site residue Asp 810.

Disregarding the standard deviations associated with the analysis, if we sum the energy contribution for each residue in the different complexes, we obtain a profile consistent with the total binding energy: **-34.23** kcal/mol for CSF-$1R^{WT}$, **-19** kcal/mol for CSF-$1R^{D802V}$, -**28** kcal/mol for KIT^{WT}, **-37.75** kcal/mol for KIT^{V560G}, **-22.80** kcal/mol for KIT^{S628N}, and **-19.58** kcal/mol for KIT^{D816V}. The energy contribution of the truncated JMR residues present at the complex KIT^{V560G} were not considered. It is clear that the summarized contribution of all residues, even minor, have an impact on the final energy of binding.

Since the charged residues play a key role in the binding energy, due to the electrostatic character of their energy contribution, we analyzed the electrostatic potential surface of the protein in order to confirm if the mutation has an effect on the charge redistribution at the protein surface.

For illustration, we have used an equilibrated conformation of the complexes, previous to the MD simulations replicas. Figure 43 confirms the hypothesis that the mutations alter the electrostatic character of the protein's electrostatic surface, especially at the vicinity of imatinib protonated region. We see clearly that in the most resistant complexes (CSF-$1R^{D802V}$ and KIT^{D816V}), the surface is less negative in comparison with CSF-$1R^{WT}$, KIT^{WT} and KIT^{V560G}. This effect could contribute to the ligand repulsion in the resistant mutants.

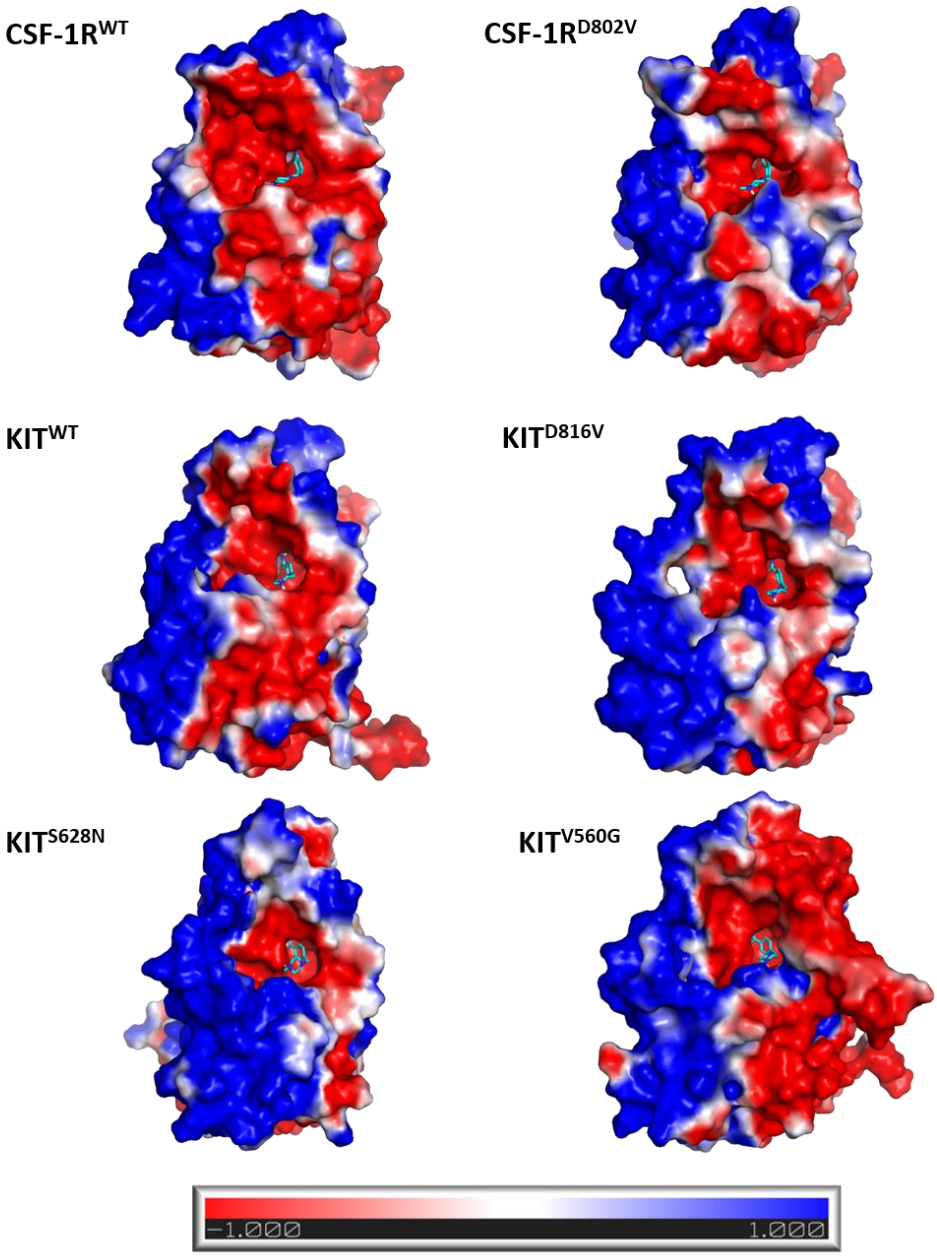


Figure 43: ***Electrostatic surface profile for all the receptor structures****. The color code corresponding to the charged nature of the surface is place in the bottom. For illustration, imatinib is placed in the ATP-binding sites and represented as sticks colored in cyan. The view of the ATP-binding site corresponds to the cavity where the protonated nitrogen is situated.*

Besides being qualitative, the electrostatic potential surface computation is not a straightforward task to be applied to the whole MD trajectory. An alternative way of measuring the variability of the electrostatic interactions in the course of MD simulations is the computation of the salt-bridges frequency. Our goal was to identify an alteration on the salt-bridges profile that could justify the presence of positive charges around imatinib's protonated portion, in the resistant mutants.

Comparing the salt-bridges profile for CSF-1R^{WT} and CSF-1R^{D802V} (Fig. 44), we observe an alteration on the salt-bridges formed the residue Asp778. As already pointed, in the WT, Asp778 is stabilizing a salt-bridge with residue Arg801, which is completely lost in the mutant. Besides the pair Asp778-Arg801, the mutant also loses the salt bridge between Asp796 and Lys616, the latter residue being identified as important contributor the final binding energy of imatinib, calculated through MM-PBSA (Fig. 41). All these residues are located close to the ATP-binding site and the presence of the positive charges no longer being stabilized by negatively charged residues could contribute to the inhibitor repulsion. In CSF-1R^{D802V}, we see that Asp778 is stabilizing Arg782 located at the C-loop, leaving Arg801 free.

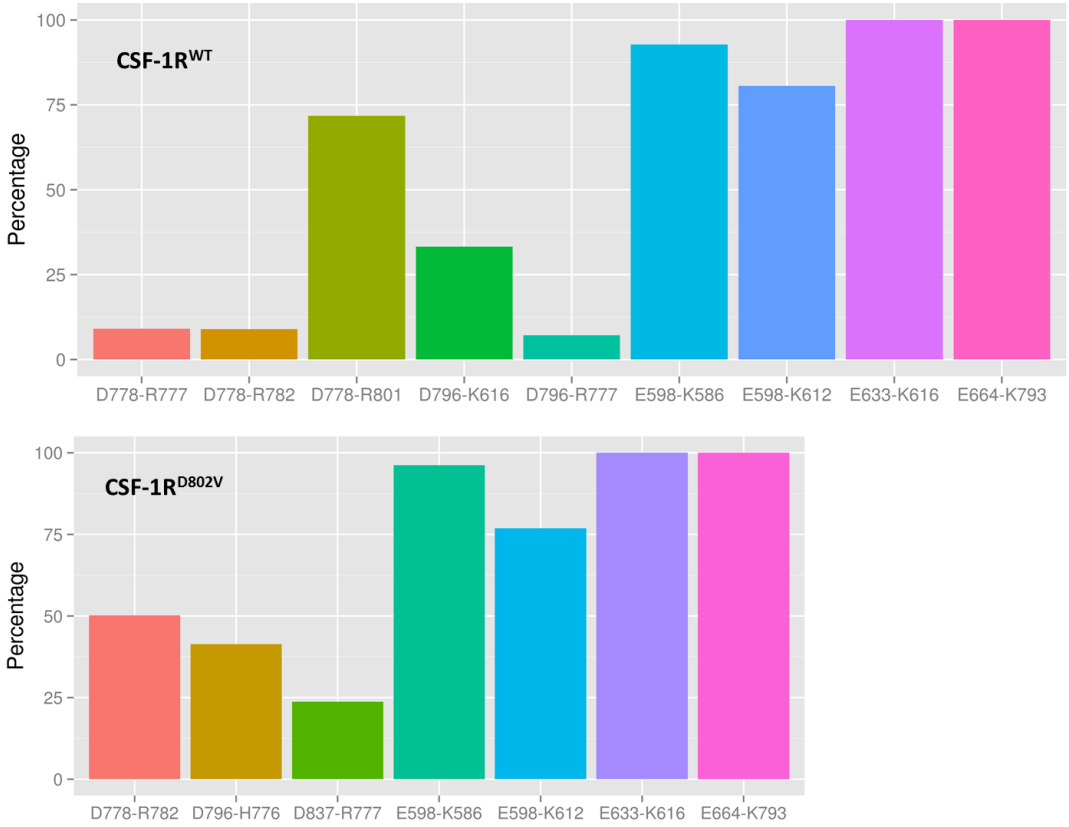


Figure 44: ***Salt-bridges profile calculated over the MD simulation replicas for WT CSF-1R and the mutant****. The occurrence of the salt bridges are represented in percentage of the MD simulations time.*

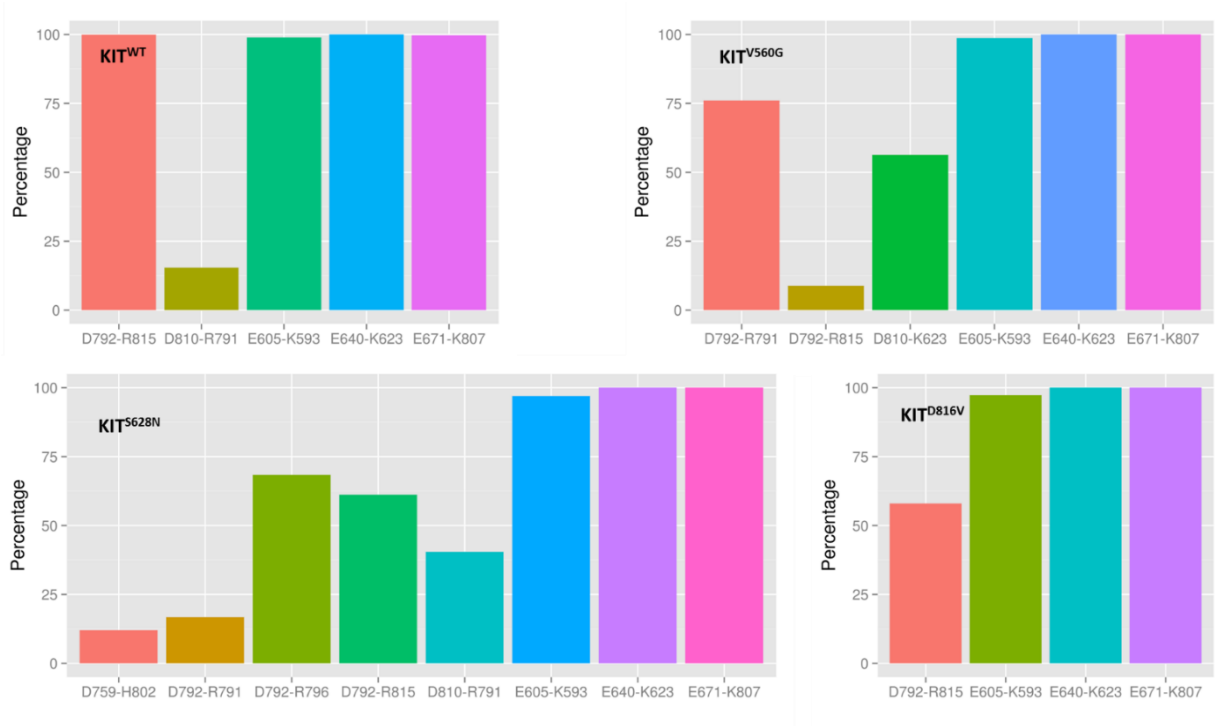


Figure 45: ***Salt-bridges profile calculated over the MD simulation replicas for WT KIT and the mutants****. The occurrence of the salt bridges are represented in percentage of the MD simulations time*

The salt-bridges profile for KIT (Fig. 45) displays more variation. Comparing KIT^{WT} and KIT^{D816V}, we observe that the mutant loses interaction for the salt-bridge formed by Asp792-Arg815, the same pair observed in the CSF-$1R^{WT}$ complex. In addition, the WT stabilizes, although with less frequency, a pair formed by Asp810-Arg791, not found in KIT^{D816V}. In the complex formed by the sensible mutant KIT^{V560G}, Asp792 stabilizes more Arg791, rather than Arg815. Interestingly it is also consistent with the energy data, where Asp792 seems to contribute more favorably to energy in comparison with the WT data. In addition, KIT^{V560G} presents an interaction Asp810-Lys623, such as found for CSF-$1R^{WT}$.

The complex formed by the mutant KIT^{S628N} is the one that shows more variation. The pair Asp792-Arg815 shows a lower frequency, when compared to the WT, but Asp792 has alternate interactions with Arg791 and Arg796, interacting more with the latter, similar to what we saw for CSF-$1R^{D802V}$. In addition, we find the pair Asp180-Arg791 with a frequency slightly higher than the values found for the WT.

The salt-bridges profiles of the studied complexes suggests that in the most resistant mutants, CSF-$1R^{D802V}$ and KIT^{D816V}, we might have lone positive charges interfering with the binding, when in the WT CSF-1R/KIT and the sensible KIT^{V560G}, these charges are stabilized

through interactions with nearby negatively charged residues. The mutant KIT^{S628N} displays an intermediate behavior between the WT KIT and the resistant mutants.

Until here, we have excluded the interference of the JMR in the binding energy, with exception of the complex KIT^{V560G}, which contains a truncated portion of this fragment and the mutation placed at the JMR fragment increases the drug sensitivity. In the previous computational studies of KIT mutants, especially KIT^{D816V}, it was evidenced that the mutation induced a long-range effect, altering the dynamical behavior of the JMR residues, which could facilitate the activation process by its departure from the TK domain (LAINE; AUCLAIR & TCHERTANOV, 2012). This long range effect is mainly observed due to the change on the net charge of the protein, caused by the mutation.

To put in evidence if the JMR has an impact on the resistance to imatinib, we have decided to repeat the MD simulations for the KIT complexes (KIT^{WT}, KIT^{S628N}, KIT^{D816V}) with the JMR fragment equivalent to the truncated portion present in the mutant KIT^{V560G}. We aimed to verify if the JMR's inclusion have an impact in the binding affinity, calculated by the MM-PBSA approach.

For a fast study-case, we have decided to analyze only the KIT complexes and we have not repeated the docking simulations. As already explained in the Methodology, we have taken the structures of KIT^{WT}, KIT^{S628N} and KIT^{D816V}, derived from the convergence analysis, excising the portion corresponding to the JMR residues 547-558, present at KIT^{V560G}. The truncated KIT structures were superposed with the final docked structures containing the low energy conformation of imatinib and the imatinib was placed manually into the ATP-binding site of the truncated structures. Possible steric clashes were eliminated by minimizing the energy of the complexes. Two MD simulation replicas (50 ns each) were produced for each imatinib-KIT form. MM-PBSA calculation was performed using the same parameters as for the previous simulations. The new results were compared with those obtained from the previous simulations of imatinib-KIT^{V560G} complex.

The comparison confirms that the length of the receptor influences the final energy values, which are lower for all studied systems in comparison with the previous calculations where the JMR was not present (Fig. 46). However, the tendency is maintained. Despite being less pronounced than the early energy calculations, the complex formed by imatinib and KIT^{V560G}

shows binding energy values lower than those found for KIT^{WT} and the difference in energy between the imatinib-bound complexes formed by the mutant KITD816V and KITWT is in the same order as in the previous calculations: 18 kcal/mol, indicating that the JMR has no direct impact on imatinib's affinity. The complex formed by KIT^{S628N} presents energy values very similar to those formed by KIT^{WT}, which goes back to the question about the sensitivity of this mutant to imatinib.

After examining the energy values, we inspected and analyzed the MD data to see if we could find other changes at the atomic level that possibly correlate with the binding energy. Figure 47 shows the RMSD calculated over two MD simulation replicas for each imatinib-KIT complex. All complexes seem to reach the stabilization before the first 10 ns, with RMSD values fluctuating around 0.15 nm for the complexes formed by KIT^{WT} and KIT^{V560G}, at 0.2 nm for the complex formed by KIT^{S628N} and 0.25 nm for those formed by KIT^{D816V}. The JMR and the C-terminal tail were excluded from the calculation, so globally the TK domain's core of the complexes maintains its stability.

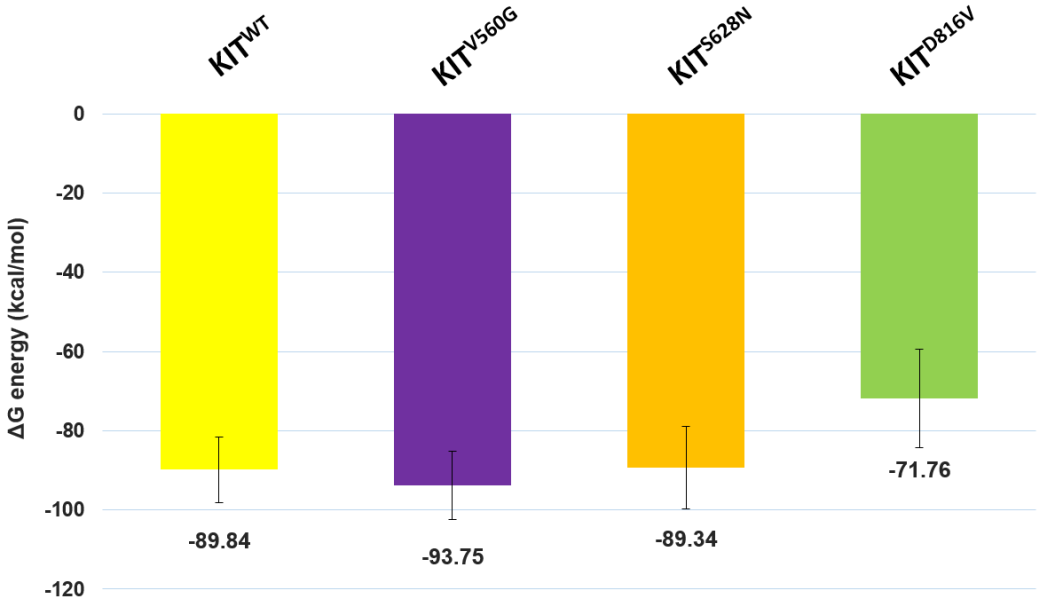


Figure 46: ***Binding energy between imatinib and the WT and mutant forms of KIT.*** *Data correspond to the new simulations test data in which all KIT forms contain the same truncated portion of the JMR, as in the complex formed by the mutant* KIT^{V560G}.

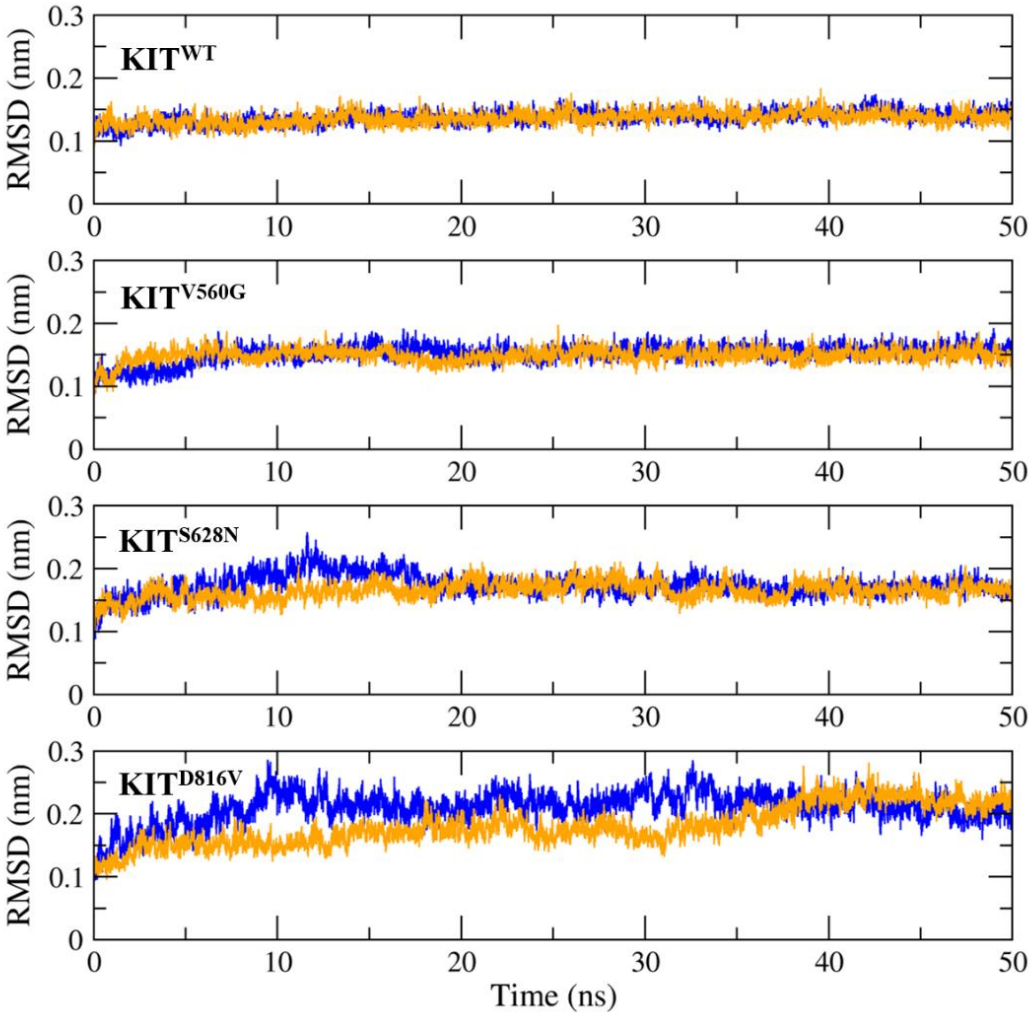


Figure 47: ***RMSD for the backbone protein atoms****. The JMR and the C-ter tail were excluded from the calculation. All the systems are labeled accordingly to each KIT form; KIT^{V560G} previous data are presented for comparison reasons. Replicas 1 and 2 are colored in blue and orange, respectively.*

Newly calculated RMSF for the protein backbone atoms and averaged from both MD replicas show that the fluctuation profile of the complexes formed by the KIT mutants is different from those formed by KIT^{WT}, in particular for KIT^{D816V} complex (Fig. 48). All complexes fluctuate very much at the truncated JMR, KID and C-terminal tail regions, which are very flexible fragments, so they were excluded from the 3D RMSF visualization (Fig. 49). The presence of the very flexible element at the N-terminal region of the TK domain (the truncated JMR) affects the stability of the entire domain, although it seems to not affect the ATP-binding site region (Fig. 49).

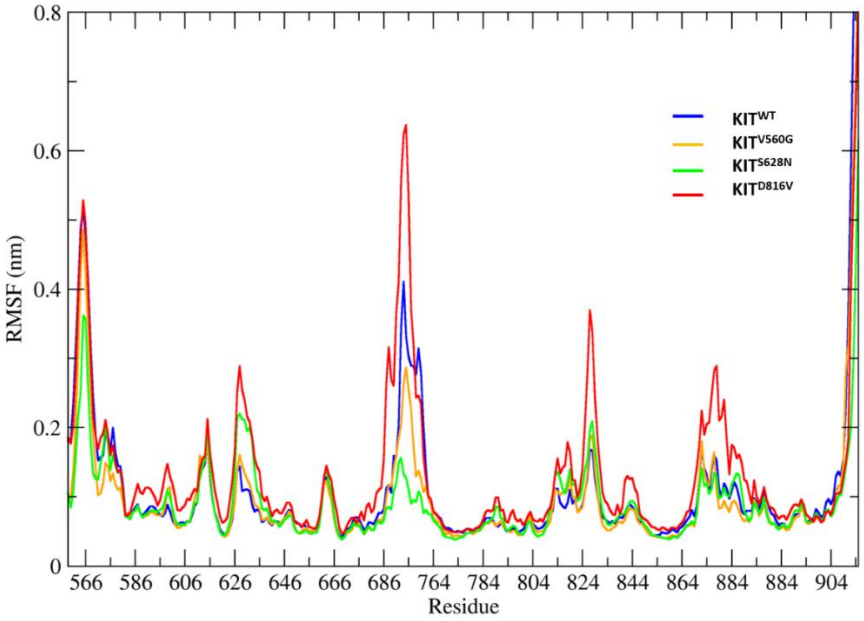


Figure 48: ***RMSF calculated for protein backbone atoms****. The different forms of KIT are label as indicated in the legend.*

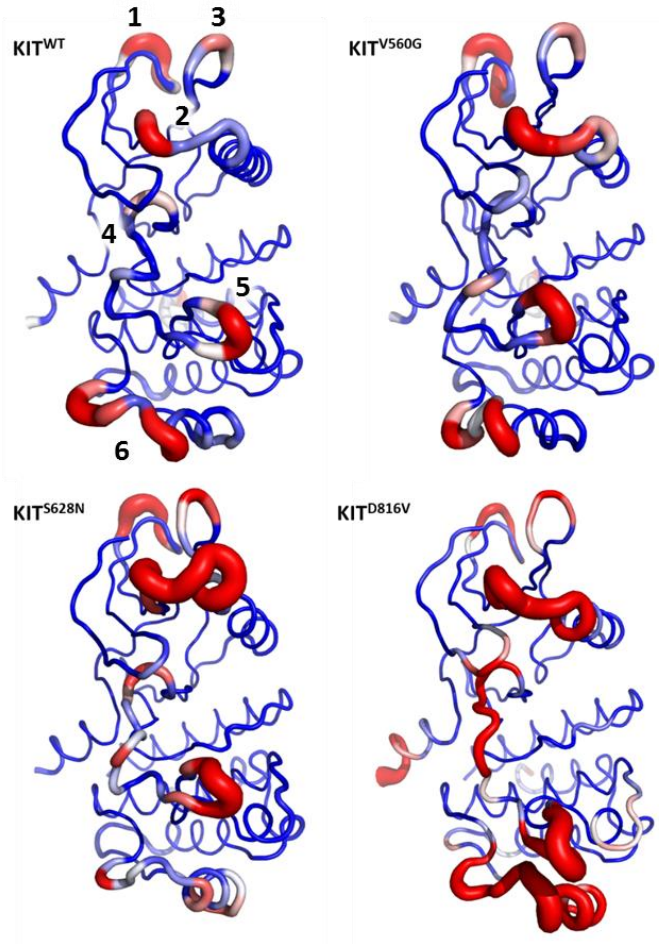


Figure 49: ***RMSF 3D representation on the protein backbone, excluding the JMR, KID and the C-ter regions****. The different conformations of KIT are labeled. The regions that fluctuate the most are thicker, colored in red and numbered in the structure of the WT.*

KIT^{D816V} presents the highest perturbation on the overall structure of the kinase domain, reaching values of 0.3 nm for the loop adjacent to the Cα-helix (**2** in Fig. 49), 0.4 nm for the A-loop`s β-harpin (**5** in Fig. 49) and 0.3 nm for the loop adjacent and α-G helix itself (**6** in Fig. 49). The residues in imatinib-KIT^{S628N} complex fluctuate most at the point of mutation, located at the loop adjacent to the Cα-helix (**2** in Fig. 49); the residues in imatinib-KIT^{V560G} complex show fluctuation values similar to those formed by KIT^{WT}.

To see in more details if the presence of the JMR affected the RMSF of the ATP-binding site region, we detailed the protein backbone RMSF for the residues located in a radius of 6 Å around imatinib. Figure 50 shows that the active site residues fluctuated with values below 0.1 nm with the exception of imatinib-KIT^{D816V} complex, whose residues can reach values of 0.2 nm. A higher fluctuation on the residues in the surroundings of imatinib could enhance the resistance effect induced by this mutation.

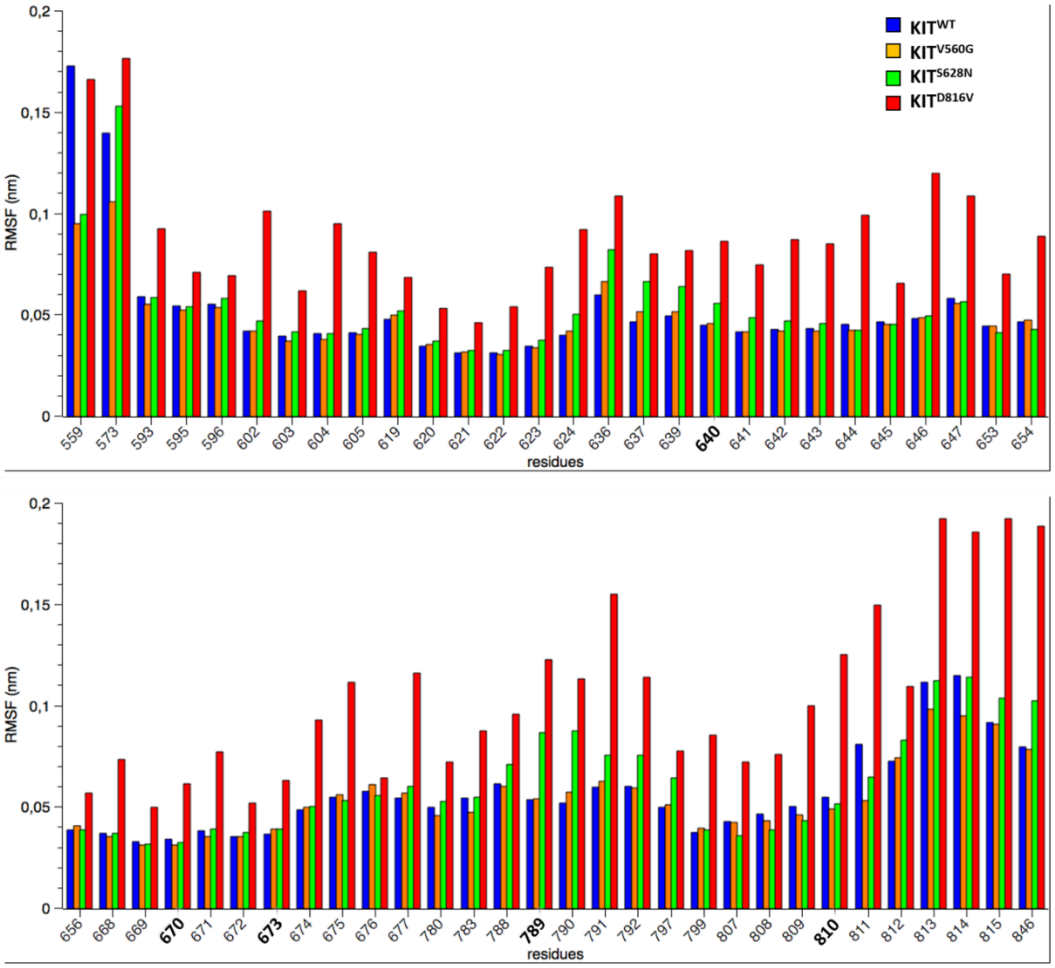


Figure 50: ***RMSF for the residues situated in a radius of 6 Å around imatinib****. RMSF values were calculated for the backbone and each bar correspond to each form of KIT. Residues that are engaged in H-bonds interactions with imatinib accordingly to the crystal structure 1T46 are highlighted in bold.*

Considering the residues that make a direct contact with imatinib, as described in the crystallographic structure of KIT complexed with the inhibitor (PDB ID: 1T46), their fluctuation values are increased in KIT^{D816V}-imatinib complex, in comparison with the complex formed by the other mutants and the WT. This is more pronounced for residues Ile789 and Asp810 whose values fluctuates around 0.14 nm against 0.1 nm for KIT^{S628N}-imatinib complex and 0.05 nm for KIT^{WT}-imatinib complex; ~0.13 nm against 0.05 nm for all the others, respectively.

Together with Cys673, residues Cys674-Asp676 belongs to the hinge region and also present a higher fluctuation in KIT^{D816V}-imatinib complex, in relation to the other systems. These residues are situated adjacent to the gatekeeper residue Thr670, very important for KIT inhibition. Thr670 controls access of the inhibitors to a hydrophobic pocket deep in the active site that is not contacted by ATP. Substitution of the gatekeeper threonine residue with bulky side chains is a common mechanism of resistance to pharmacological ATP-competitive kinase inhibitors (AZAM *et al.*, 2008).

The first imatinib-resistant mutation described in CML patients was an isoleucine substitution at the gatekeeper residue threonine (GORRE *et al.*, 2001), reinforcing the importance of this residues and the contact between the inhibitor and the other hinge residues. Residues from the catalytic loop, such as Arg791 and Asp792, and the A-loop residues, Phe811-Arg815, also fluctuates most at KIT^{D816V} complex, probably due to the local impact of D816V mutation. Despite the increased fluctuation in the residues mentioned above, for the KIT^{D816V} complex, the H-bond pattern it is not dramatically changed (Tab. 10) when compared with the previous MD simulations, in absence of the truncated JMR (Tab. 7).

The complex formed by KIT^{WT} shows slightly different prevalence values in both simulations, especially concerning residues Asp810 and Glu640, with increased values and a diminution on the bond prevalence of Thr670. Complexes formed by KIT^{S628N} and KIT^{D816V} still make highly prevalent H-bonds with Cys673, but now both mutants display the profile of interaction with Asp810 seen previously only for the mutant KIT^{S628N}, in which the inhibitor makes alternate interactions with the backbone nitrogen of Asp810 and its δ-oxygens (Tab. 10).

Table 10: Hydrogen bonds (H-bonds) occurrences between the targets and imatinib, averaged over the two MD replicas. The atom pairs for donor and acceptor interactions are depicted in the table. Imatinib atoms participating in the interaction are represented in figure 30.

	H-bond occurence (%)					
	N3-(N)Cys673	**O1-(N)Asp810**	**N7-(Oδ)Asp810**	**N7-(O)Ile789**	**N5-(Oε)Glu640**	**N4-(Oγ)Thr670**
KIT^{WT}	99	73	0	99	31	17
KIT^{V560G}	98	73	0	96	47	35
KIT^{S628N}	99	60	81	0	60	25
KIT^{D816V}	98	49	48	70	48	1

Visual inspection of the trajectory confirmed the flipping of the methylpiperazin in one of the MD simulations of KIT^{D816V} complex. Moreover, in KIT^{S628N} complex, the side chain of Ile789 suffers a dramatic change during the simulations, which abolish the interaction of imatinib with this residue (Tab. 10). The same is observed in KIT^{D816V} complex but these changes oscillated a lot during the MD replicas.

Another interesting feature is the complete loss of H-bond between imatinib and the gatekeeper residue Thr670 is in KIT^{D816V} complex, probably due to the increased fluctuation of the hinge region in this mutant. A 2D-representation of the H-bond profile concatenating the information of the previous and current simulations is shown at Figure 51.

The new data from MD simulations of KIT in presence of the truncated JMR suggests that the mutation induced effects on the JMR structure, probably by allosteric propagation, can contribute to destabilize the inhibitor inside the ATP-binding site of the resistant mutants, specially interfering with the KIT^{D816V} complex.

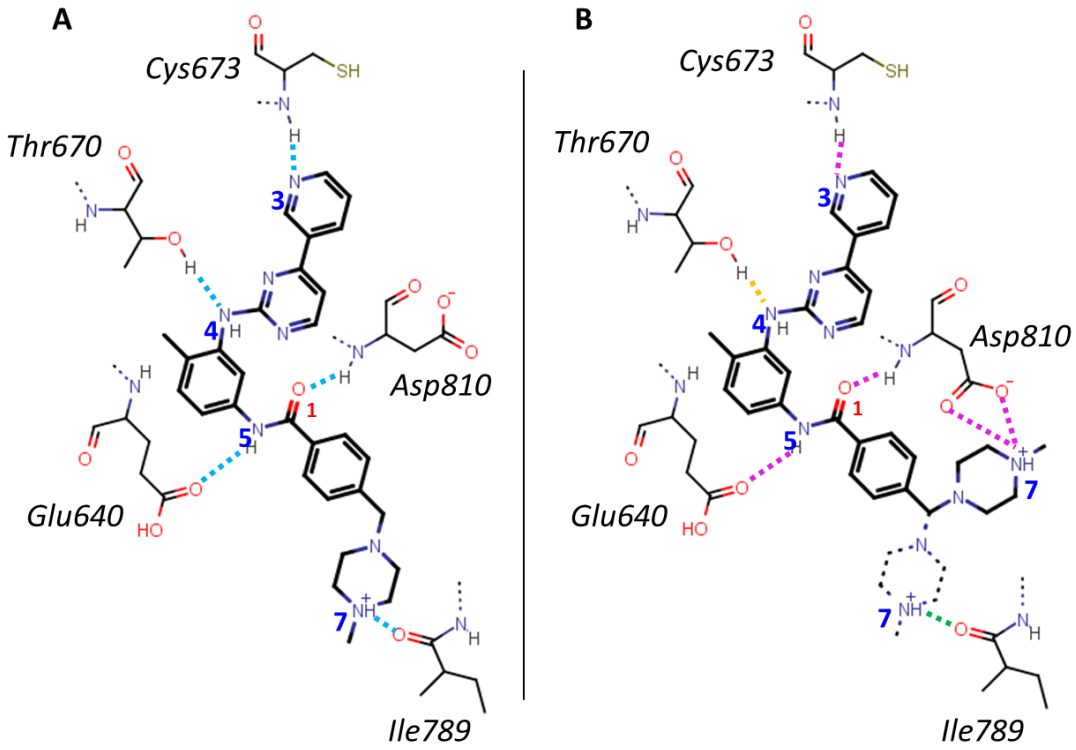


Figure 51: ***2D representation of imatinib and the ATP-binding site residues involving in H-bond interactions with the inhibitor.*** *(A) H-bonds pattern concerning the systems KIT^{WT} and KIT^{V560G}. We see that the side chain orientation of the protein residues remains intact. (B). H-bonds concerning the systems KIT^{S628N} and KIT^{D816V}.In both systems, the side chain orientation of Asp810 changes, which facilitates the H-bond between AspOδ and N7 from imatinib. The bonds coloured in green and orange represent the contacts that were lost in the second run of MD simulations, for KIT^{S628N} and KIT^{D816V}, respectively.*

Discussion

The conformational plasticity of RTKs endows these receptors with a wide range of functions that must be tightly tuned. Gain-of-function mutations can alter this fine-tuning at different levels, including ligand binding, receptor dimerization, kinase domain conformation transition, and post-translational modifications. All these modifications lead to the abnormal signalization of the concerned cells, in the case of cancer, increasing the cell proliferation (VERSTRAETE & SAVVIDES, 2012).

As a secondary effect, these mutations can also trigger the sensitivity, as demonstrated for the V560G substitution in KIT (FROST *et al.*, 2002), or resistance to tyrosine kinase inhibitors such as imatinib, as demonstrated for the substitutions S628N (VITA *et al.*, 2014) and D816V in KIT (FROST *et al.*, 2002), and the not-frequently found mutation D802V in CSF-1R (TAYLOR *et al.*, 2006).

The first part of this thesis was dedicated to investigate the structural and dynamical effects of the D802V mutation in the CSF-1R, and to relate our results with the ones obtained previously for KIT mutations (CHAUVOT DE BEAUCHÊNE *et al.*, 2014; DA SILVA FIGUEIREDO CELESTINO GOMES *et al.*, 2014; LAINE *et al.*, 2011; VITA *et al.*, 2014).

It is well stablished that the JMR coupling with the TK domain controls the receptor activation process. The phosphorylation of residues Tyr568 and Tyr570 within JMR induces the detachment of JMR from the kinase C-lobe and increases the fluctuation in the structure of JMR, thus appearing to initiate the kinase activation process (ZOU *et al.*, 2008). Studied by our group in ENS Cachan, the *in silico* characterization of the D816V mutation in KIT evidenced that the mutation caused local impact on the A-loop structure, followed by a structure reorganization of the JMR which facilitated its departure from the TK domain (LAINE *et al.*, 2011).

By combining various methods to analyze and compare the structure and dynamics of the native and mutated KIT and CSF-1R, the present study demonstrated that the two homologous mutations, D802V and D816V, do not have the same consequences in terms of receptor conformation and dynamics.

The local impact of D802V mutation, which is a partial unfolding of the small 3_{10}-helix (*H2*) at proximity of the mutation site in CSF-1R, is very similar to that observed in KIT D816V (LAINE *et al.*, 2011). However, we could not observe any tendency of departure for the JMR domain. In KIT simulations, the departure conformation of JMR domain, derived from the crystal 1T45 (MOL *et al.*, 2004) was mainly unstructured and gained a fold in beta-hairpin only for the D816V mutant. In the contrary of KIT, CSF-1R's native initial configuration, derived from the crystal 2OGV (WALTER *et al.*, 2007) was already well-folded, containing the beta-hairpin in the JM-S region, such as for the structure seen in KIT simulations for the D816V mutant (LAINE *et al.*, 2011).

The strong coupling of the JMR to the TK domain in CSF-$1R^{WT}$ and mutant was confirmed by PCA, normal modes and hydrogen-bonds calculations. The latter have shown that the JM-S interacts extensively with the TK domain of the receptor, which could prevent its departure during the MD simulations. Another factor is the simulated time of the trajectories, which could not be sufficient to see such effect. In addition to that, the JMR of KIT and CSF-1R present differences in their primary sequence, as seen in the Figure 26, which justify a strong complementarity between the JMR and the TK domain.

It has been described previously that allosteric coupling can be mediated solely by transmitted changes in the protein dynamics/motions as a consequence of a re-distribution of the protein conformational populations (CHENNUBHOTLA & BAHAR, 2006, 2007; CHENNUBHOTLA; YANG & BAHAR, 2008; PANDINI *et al.*, 2012; PIAZZA & SANEJOUAND, 2009).

The analysis of the CPs in the cytoplasmic domain of CSF-1R and KIT (LAINE; AUCLAIR & TCHERTANOV, 2012), showed that the two mutations, D802V and D816V, disrupt the allosteric communication between two essential regulatory fragments of the receptor, the JMR and the A-loop. Nevertheless, the disruption in CSF-1R is only partial, since the JM-B portion of the JMR maintains communication with the TK domain. Another interesting result derived from MONETA was the difference between the CPs between WT CSF-1R and KIT, being much stronger in the first case. So it would be more difficult to disrupt the strong protein network in CSF-1R.

The differential impact on the conformational dynamics of both receptors was also related with differences in the 3D structures of activated CSF-1R (PDB ID: 3LCD) and KIT (PDB

ID: 1PKG). The superposition evidenced that the activation of KIT provokes a bigger conformational change in the cytoplasmic domain of the receptor, in comparison with CSF-1R (Fig. 29). The experimental attempts of expressing the mutant receptor CSF-1R bearing the D802V mutation were only possible in the presence of a second concomitant mutation in a tyrosine located at the KID region (Y708F). This substitution was shown to enhance the receptor stability in the membrane (MORLEY *et al.*, 1999). Allied to the subtle effect of this mutation in the dynamics of CSF-1R, as seen in this study, this could explain why the D802V mutation is not frequently found in cancer.

Our data allowed us to hypothesize that the subtle effect of this mutation in the protein dynamics might be sufficient to stabilize an intermediate active conformation of the mutant CSF-1R receptor, since the N-ter portion of the JMR loses interaction with the TK domain, as seen by H-bonds and MONETA analysis, and we observe briefly a conformational change in the side chain of the phenylalanine present in the DFG motif (Fig. 25D). This conformation might not accommodate imatinib, leading to the resistance.

The second part of this thesis was dedicated to investigate the sensitivity/resistance of the WT and mutant forms of CSF-1R and KIT. Despite being insightful, the study of the receptors in the presence of imatinib is necessary to understand and complement the previous works involving both receptors in their isolated forms.

All the complexes showed good docking results, reproducing the crystallographic pose of imatinib and were very stable during the MD simulations. A new interaction with imatinib was found for the C-loop residue Ile775/789. This interaction suffers a dramatic loss for the complex formed by CSF-1R^{D802V}, probably due to the increased flexibility of this loop in the MD simulations for this mutant. However, excepting KIT^{S628N}, all KIT complexes showed very stable H-bond interactions values, even for the most resistant mutant, KIT^{D816V}.

Despite interacting similarly at the MD level, WT and mutants showed binding energy values with the same tendency observed for the experimental data of inhibition obtained for imatinib. The electrostatic interactions between the protonated inhibitor and the negatively charged residues in the ATP-binding site vicinity showed to be the main factor that drives the sensitivity or the resistance to imatinib.

The entalpic energy decomposition by protein residue has highlighted that the Asp to Val substitution in CSF-1R^{D802V} and KIT^{D816V} was the most powerful in losing of energy contribution to the final binding energy values. For the other KIT mutants, the mutated residues had no significant difference in energy contribution, compared with their equivalent in the WT KIT (Tables 12 and 13). The total sum, however, showed an importance difference in the ΔG, consistent with the difference observed for the final relative binding energy values. This suggests a global change in the electrostatic profile of the protein, not being as localized as for D802V and D816V mutations.

Indeed, normalizing the residue contribution by subtracting the values obtained for KIT^{WT} (Fig. 42), we see that KIT^{V560G} has more residues contributing favorably to the binding energy, even extracting the contribution of the extra residues present on the structure due to the truncated JMR. For KIT^{S628N} the difference is less clear, which makes us question the role of this mutation. By the experimental data concerning the S628N mutant, we see that this mutant is more resistant than the WT KIT, however it is more sensible than the well-characterized resistant KIT^{D816V} (VITA *et al.*, 2014). The instability of the ligand inside the ATP-binding site of the S628N mutant, as evidenced by its varied H-bond profile, could explain the loss of sensibility to the drug.

In an attempt to account for the electrostatic interactions during the MD simulations, we calculated the salt-bridges profile for all the complexes. The results suggests that in the resistant D802V and D816V mutants, there is a charge redistribution in the vicinity of the ATP-binding site that could favor the ligand repulsion. This is less pronounced for the S628N mutant.

Aleksandrov and his group described the probable protonation states of imatinib in solution and in complex with a kinase protein (ALEKSANDROV & SIMONSON, 2010). According to their free energy studies, imatinib binds to Abl in its protonated, positively charged form. Although, it is also mentioned, referring to another publication (SZAKÁCS *et al.*, 2005), that the inhibitor in solution at a pH of 7.4 spends 2/3 of its time in a neutral state. It should be interesting to study the inhibitor in its deprotonated form to see if the results are dissimilar concerning WT and the D802/816V mutants, since the electrostatic profile of the protein is changed upon mutation.

We decided to test the influence of the JMR on the binding energy, since the mutant KIT^{V560G} contains a truncated portion of this fragment, which probably originated the very low binding energy value for this complex. Our data indicates that the JMR has a minor role in the sensitivity/resistance mechanism, since the sensible mutant still have lower energy values than the WT and the resistant D816V has the ΔG in the same order of difference than observed before (~18 kcal/mol) (Fig. 46). The energy values for S628N are very similar to those found for the WT, reinforcing the questionable resistance of this mutant. The MD data has also pointed to a possible destabilization of the ATP-binding site residues, especially for the D816V mutant, which is consistent with the allosteric effect propagated by this mutation (LAINE; AUCLAIR & TCHERTANOV, 2012).

Conclusions

The present study has demonstrated that the two homologue mutations, D802V in CSF-1R and D816V in KIT, do not have the same effect on the receptors structure and dynamics. The two mutations have a local impact in the A-loop structure and disrupt the allosteric communication between the A-loop and the JMR. Nevertheless, the disruption in CSF-1R is not sufficient to induce the JMR's departure from the TK domain, due to the strong coupling between the JMR's distal region and the TK domain.

This differential impact on the conformational dynamics of the receptor was related to differences in the primary sequence in the JMR between the two wild-type receptors. The partial loss of interactions of the JMR with the TK domain could be sufficient to stabilize an intermediate active conformation of the A-loop unfavorable for the binding of imatinib. In addition, the subtle effect of the D802V mutation could explain the low incidence of this mutation observed in clinic.

A better understanding of the sensitivity/resistance mechanism to imatinib by the presence of oncogenic mutations was complemented by the study of the receptors in the presence of the inhibitor. The energy of binding between imatinib and the different targets was coherent with the experimental data, showing that the MM-PBSA approach is valid to estimate the relative order of the binding energies. Altogether, the decomposition of the binding energy into the different terms that contribute to the final energy values, followed by the energy contribution of each protein residue and the salt-bridges profile pointed to the electrostatic interactions as the main factor determining the affinity of the targets to imatinib. The JMR does not seem to interfere significantly with the binding of the inhibitor, having a minor role in the resistance mechanism.

The *in silico* approaches applied on this thesis have shed light on the oncogenic activation mechanism of the CSF-1R receptor, expected to be similar to the homologue KIT mutant. In addition, the study of KIT and CSF-1R in their WT and mutant forms in complex with imatinib have complemented the early structural studies of the isolated systems, showing that not only the conformational change associated with the activation, but also the electrostatic interactions and the protonation state of the ligand can explain the resistance phenomena.

Appendix

1. Modeling of the full length structure for the CSF-1R`s cytoplasmic domain by prediction of the KID structure

Modeling CSF-1R´s full length structure is of interest as a perspective for this thesis. The alternative phosphorylation observed experimentally for KID´s tyrosines is associated with a different distribution of the receptor onto the cells membrane (unpublished data). Due to the preliminary character of this data, we decided to present it as an appendix.

1.1. Methodology

1.1.1. Bioinformatics analysis

Before modeling the KID, we have submitted the KID sequence (residues 680-751) to BLAST (ALTSCHUL *et al.*, 1990), using the following parameters: blastp algorithm, the PDB as the search database and the organism set to human. Structures derived from comparative models were excluded from the search. Unfortunately, BLAST has not returned any viable candidate template to be used in a comparative modeling attempt, with exception of one crystallographic structure of CSF-1R containing a small portion of the KID solved (PDB ID: 3LCO). The portion of KID sequence covered by this structure corresponded to residues 680-686,747-751. In addition, we compared the KID sequence of CSF-1R with the other members of type III RTK family using CLUSTAWL(MCWILLIAM *et al.*, 2013) .

The secondary structure prediction for the KID sequence, flanked by 8 extra residues at each extremity (residues 672-759), was performed using different methods: GOR4 (GARNIER; GIBRAT & ROBSON, 1996), SOPMA (GEOURJON & DELÉAGE, 1995), SSPRO (POLLASTRI *et al.*, 2002), PORTER (POLLASTRI & MCLYSAGHT, 2005), SAM_T08 (using the alphabet dssp_ehl2) (KARPLUS, 2009), PSIPRED (MCGUFFIN; BRYSON & JONES, 2000) and JPRED (COLE; BARBER & BARTON, 2008).

In addition to the secondary structure prediction, we have submitted the KID sequence to the IUPRED server, which presents an algorithm for predicting intrinsically unstructured/disordered proteins and domains (IUPs) from protein sequences by estimating

their total pairwise interresidue interaction energy. The method is based on the assumption that IUP sequences do not fold due to their inability to form sufficient stabilizing interresidue interactions (DOSZTÁNYI *et al.*, 2005). More details about the method can be found at the publication. We have selected the option *short disorder*, recommended for missing regions in crystallographic structures.

1.1.2. Modeling protocols

The modeling of the kinase insert domain (KID) was performed according to two protocols, described in the next subsections.

Protocol 1

The following procedure was reproduced from the protocol made by Isaure de Beauchêne at her PhD thesis (CHAUVOT DE BEAUCHÊNE, 2013). Briefly, it consists of using Rosetta (ROHL *et al.*, 2004) to perform *de novo* prediction of the structure based on a sequence containing the KID flanked by 5 extra residues at each extremity of the KID sequence (KID+10: residues 675-756). After the KID prediction, the chosen structure was inserted into the TK domain using Modeller (ESWAR *et al.*, 2008).

The module *AbinitioRelax*, available with Rosetta (2014 version), was used to perform the *de novo* prediction of 10.000 models for the KID+10 sequence. Besides the target sequence in FASTA format, this module requires a 3-residues and a 9-residues sequence-specific fragment files, that can be generated at the web-server Robetta (KIM; CHIVIAN & BAKER, 2004), and an optional secondary structure prediction file from PsiPred (MCGUFFIN; BRYSON & JONES, 2000). The Robetta server uses Rosetta to generate fragment libraries and also 3D protein structure models *de novo*, but the output data are restricted to 5 best predicted models.

The *AbinitioRelax* application consists of two main steps. The first step is a coarse-grained fragment-based search through conformational space using a knowledge-based "centroid" score function that favors protein-like features (*Abinitio*). The second optional step is all-atom refinement using the Rosetta full-atom force field (*Relax*) (BRADLEY; MISURA & BAKER, 2005).

The 10.000 generated models were clustered using the module *cluster* available with Rosetta. The *cluster* application algorithm finds the structure with the largest number of neighbors within the cluster radius and creates a first cluster with that structure as the cluster

center and the neighbors are part of and claimed by the cluster. The structures are then gradually removed from the pool of "unclaimed" structures and the algorithm is repeated until all structures are assigned a cluster. The models generated by Rosetta were very different from each other in structure and the algorithm was incapable of grouping more than one structure in a cluster, so we decided to sort the structures according to their energy.

The 100 lower energy models were analyzed by an in-house script that verified if the distances between the N and C-terminal extremities of the models were in accordance with a specific value (13 Å), allowing a deviation of 2 Å maximum from the specific value. This value was extracted from the distance between the two residues located at the extremity of the KID+10 sequence, according to their position in the crystal structure of auto-inhibited CSF-1R (PDB ID: 2OGV, residues 675 and 756) (WALTER *et al.*, 2007). This step was necessary to assure that the KID predicted structure would be viable to be inserted into the TK domain using Modeller.

Six final KID structures corresponded to the distance criteria (Tab. A1). For each one of the six Rosetta models, 100 models were generated using Modeller in order to insert the KID into the TK domain. The insertion of the predicted KIDs into the TK domain was performed by the multiple templates approach, using as template for the TK domain the crystal structure of auto-inhibited CSF-1R (PDB ID: 2OGV) and the predicted KID structures for the KID region.

Table A 1: Models issued from the Abinitio.relax application, selected by the distance between their N- and C-terminal residues. Models are ranked accordingly to their energy. Model ID can range from 1 to 10.000. The value in parentheses correspond to the deviation amount in relation to the specified distance value of 13 Å.

Model ID	Distance between terminals (Å)	Rosetta energy score
5494	13.29 (0.14)	-62.92
2024	14.37 (1.23)	-61.58
2131	14.72 (1.58)	-61.22
6393	12.82 (0.32)	-61.04
7693	14.04 (0.90)	-60.95
2518	13.98 (0.84)	-60.42

Before constructing the models, we have manually altered the alignment between the templates by deleting the residues corresponding to the truncated KID present at crystal 2OGV. The deletion was also done at the PDB structure. This was necessary to assure that Modeller would only consider the predicted KID as template for the KID region. Best models issued from each Modeller essay can be found at Table A2.

Table A 2: Best models generated by Modeller during the independent essays associated with each one of the six models selected from Rosetta. Each Modeller essay was composed of 100 runs and the Table is ranked by the best models from each essay, based on their DOPE score.

Rosetta model ID	Modeller model ID	DOPE score
5494	69	-40679.13
2024	26	-40536.40
2518	7	-40344.64
2131	69	-40120.19
6393	1	-39944.61
7693	53	-39527.67

Protocol 2

In this protocol, we decided to couple the homology modeling with *de novo* prediction. Predicting the KID tridimensional structure only by comparative modeling is not viable, since there is no homologue structures to use as templates for the modeling. However, we decided to use as a template for the KID, the structure of CSF-1R (PDB ID: 3LCO) that contains a small solved region of the KID, found by a BLAST search described earlier.

We have performed a multiple-template comparative modeling using Modeller, having the auto-inhibited structure of CSF-1R (PDB ID: 20GV) (WALTER *et al.*, 2007) as template for the TK domain and the crystal structure with a partially solved KID (PDB ID: 3LCO) to be considered as template only for the KID. For this, we had to do the same procedure as in the preceding protocol: before aligning the full-length cytoplasmic domain sequence of CSF-1R with the two templates, we had to manually align the templates sequences corresponding to the PDB structures 2OGV and 3LCO and delete the residues corresponding to the pseudo-KID, present at 2OGV, from the sequence and the 2OGV`s PDB.

One hundred (100) models were generated using the default parameters of Modeller. All models contained the small helix portion present at the structure 3LCO and the rest of the KID

was mainly disordered. The structure containing the best DOPE score was selected to the phase 2 of this protocol.

In the phase 2, we have used Rosetta's module *loopmodel.default* to generate structures of the full protein by performing the *de novo* reconstruction only of the remaining KID region that was absent from 3LCO, that is the disordered region that Modeller failed to model since there were no templates available (residues 685-748). The loop modeling was performed with the parameters from the *kinematic loop modeling,* in order to reconstruct the structure of the desired sequence. The program discards the initial backbone and side-chain conformations. Due to the very long nature of Rosetta's calculations, only 431 models were generated.

The remodeling is signaled to Rosetta by adding the flags –loops:remodel perturb_kic. Kinematic closure (KIC) is an analytic calculation inspired by robotics techniques for rapidly determining the possible conformations of linked objects subject to constraints. *2N - 6* backbone torsions of an *N*-residue peptide segment (called non-pivot torsions) are set to values drawn randomly from the Ramachandran space of each residue type, and the remaining 6 phi/psi torsions (called pivot torsions) are solved analytically by KIC. This formulation allows for rapid sampling of large conformational spaces (MANDELL; COUTSIAS & KORTEMME, 2009). The generated models were clustered using the module *cluster*, available with Rosetta. The top seven best models from each cluster were retrieved from the output and their scores can be seen at Table A 3.

Table A 3: Cluster analysis performed on the generated models from the loopmodel application. Structures were grouped into seven clusters, with the cluster ID ranging from O to 6. In the Table, the clusters are ranked by the score of the top seven best models, indicated by their Model IDs, that ranges from 1 to 431.

Cluster ID	Cluster size	Score for the best models	Model ID
6	32	-612.923	221
1	67	-609.541	293
5	19	-605.902	346
0	255	-602.048	54
3	24	-592.163	276
2	18	-597.15	378
4	16	-591.27	299

As an alternative to Rosetta, we have decided to combine Modeller prediction with another software that makes *de novo* prediction for the refinement/optimization of loops. We decided to use CABS-FOLD web-server (BLASZCZYK *et al.*, 2013) due to a recent study that compared different modeling techniques in order to stablish the best multi-method approach for the modeling of loops (JAMROZ & KOLINSKI, 2010). The method represents the protein as a coarse-grained model of protein chains and uses statistical potentials derived from known structures. The exploration of the conformational space is done by a Replica Exchange Monte Carlo scheme with 20 replicas spanning the specified temperature range, pre-defined by the web-server (BLASZCZYK *et al.*, 2013).

In practice, we submitted a single template to the web-server with a missing fragment, in our case, the same region of the KID used in the loop reconstruction module of ROSETTA (residues 685-748). The models were generated using the default parameters.

Among the outputs provided by CABS-fold server, one can find a trajectory in Cα representation containing 400 frames for the folding procedure, as well as PDB files for the predicted models. The predicted models correspond to each cluster representative structure, e.g. the model which average dissimilarity to all models in a cluster is minimal. The clusters are numbered accordingly to their density, from the densest to the least dense one (Table A4).

Table A 4: Clustering data obtained from the CABS-fold run using consensus modeling default temperature parameters. The clusters are ranked according to their density, from the most populated to the least populated one.

Cluster ID	Cluster size	Cluster RMSD (Å)
1	75	2.2
2	65	2.6
3	63	3.2
4	38	4.1
5	27	3.4
6	27	4.2
7	24	4.4
8	21	4.7
9	13	4.6
10	7	7.2

1.2. Preliminary results

As already mentioned in the Introduction, the kinase insert domain (KID) is a very flexible region of the TK domain, generally not fully solved or truncated in the crystallographic structures. Since there is no viable templates to be used in comparative modeling runs, the prediction of the full-length CSF-1R cytoplasmic domain (CSF-1R^{full}) was performed via a combination of *de novo* and comparative modeling techniques. The *ab initio* and *de novo* techniques are capable of predicting the tridimensional structure from its amino-acid sequence.

We first submitted the KID sequence (residues 680-751) (Fig. A1) to BLAST (ALTSCHUL *et al.*, 1990), using the *blastp* algorithm, in order to confirm the absence of similar solved structures with similar sequences. As expected, it returned no viable candidate, except for a crystal structure of CSF-1R containing a small portion solved for the KID (PDB ID: 3LCO), covering a region corresponding to residues 680-686,747-751 (Fig. A 2).

```
        680                                                                  751
         |                                                                    |
LNFLRRKAEAMLGPSLSPGQDPEGGVDYKNIHLEKKYVRRDSGFSSQGVDTYVEMRPVSTSSNDSFSEQDLDKEDGRPLELRDLLHFS
```

Figure A 1: ***Primary sequence of the kinase insert domain (KID)****. In black, is represented the real sequence of the KID and in green and red, respectively, are the extra residues used in the modeling protocols and the secondary structure analysis.*

Next, we went to verify if the KID sequence is variable among the type-III RTK family. Aligning individually CSF-1R with each member of the family, we observe that the sequence is very variable in number and residue nature (Fig. A 3).

We have equally performed the secondary structure prediction analysis for the KID sequence, flanked by eight extra residues from the TK domain (Fig. A 1). Different methods were used but the consensus found the structure predominantly disordered, with the most part of the sequence being in coil (Fig. A 4).

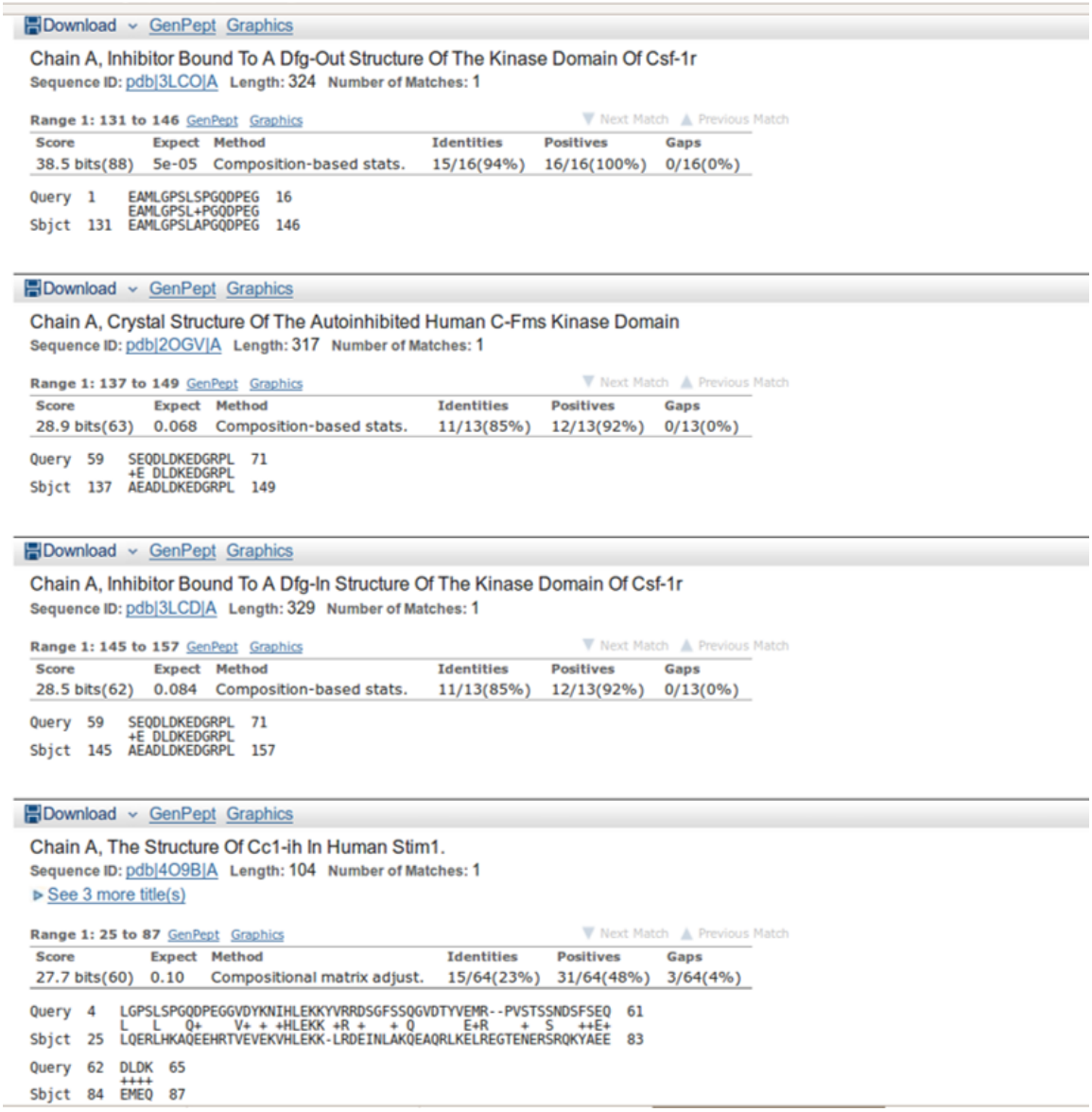


Figure A 2: ***Blast most significant results.*** *The crystal structure of CSF-1R 3LCO is the one that has the best coverage for the KID sequence, containing the residues 680-686, 747-751.*

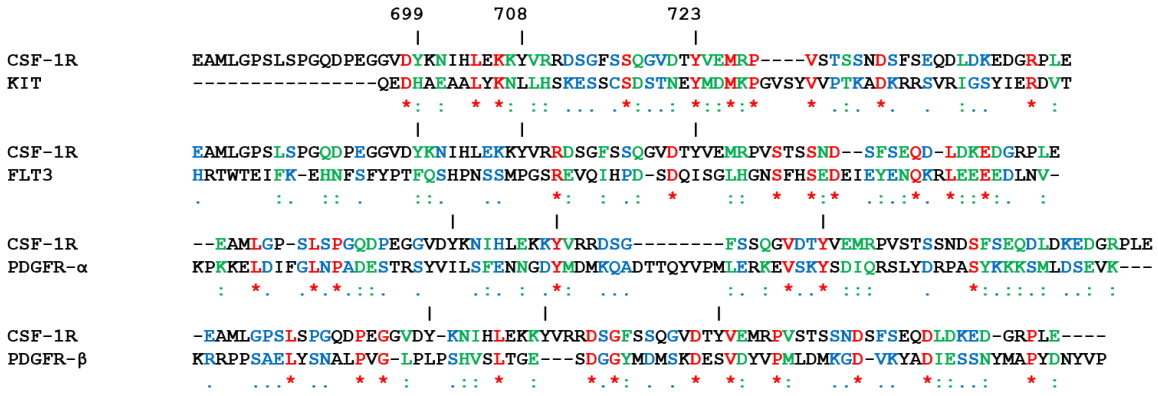


Figure A 3: ***Comparison between the sequences for the KID region among all the members of type III RTK family****. Conserved residues (*), conservative mutations (:) and semi-conservative mutations (.) are colored in red, green and blue, respectively.*

To confirm the disordered nature of CSF-1R's KID, we have submitted the sequence to the IUPRED server (DOSZTÁNYI *et al.*, 2005), which is capable of predicting intrinsically unstructured proteins and domains. Figure A 5 shows that among the 72 total residues of the KID, 69 residues have values equal or superior to 0.4 and 43 of them have values higher than 0.5, which is the indicative of disorder predicted by the server.

```
                672       680       690       700       710       720       730       740       750       759
                  |         |         |         |         |         |         |         |         |         |
                  LNFLRRKAEAMLGPSLSPGQDPEGGVDYKNIHLEKKYVRRDSGFSSQGVDTYVEMRPVSTSSNDSFSEQDLDKEDGRPLELRDLLHFS
GOR4              ccccccchhhhccccccccccccccchhhhhhhhceecccccccccccceeeeeeccccccccchhhhhhhccccchhhhhceeeeec
SOPMA             hhhhhhhhhheeccccccccccccchhhhhhhhheeccccccccccceeeeeccccccccchhhhhccccccchhhhhhhhhc
SSPRO             cchhhhhhhhhhccccccccccccccccceeeeeeeeeccccccccccceeeeccccccccccchhhhhhccccccchhhhhhcc
PORTER            cchhhhhhhhhhccccccccccccccccccccccchcccccccccccccecccccccccccccccccccccccchhhhhhhc
SAM_T08           hhhhhhhhhhhccccccccccccccccccccecceeeeccccccccccccheeeccccccccccccccccchccccccchhhhhccc
PSIPRED           cchhhhhhhhhhccccccccccccccccccccccccccccccccccccccceeeecccccccccccccccccccccccccccccc

Consensus         cchhhhhhhhhhccccccccccccccccccc----eecccccccccccceeeeccccccccccc---cchccccccchhhhhh-c

Burial            AAAEDADECDFEDDBEBAAAAAAAAABEDEFBDBBBEFDDFFFFFDBADABEEDFEBEBAAAAAADAAAAAAAAAABDEBFDEFFBDA

e = strand
h = helix
c = coil

Burial index
A         count < 27
B         27 <= count < 34
C         34 <= count < 40
D         40 <= count < 47
E         47 <= count < 55
F         55 <= count < 66
G         count >= 66
```

Figure A 4: ***Secondary structure prediction for CSF-1R's KID using different methods****. The consensus prediction was done manually; regions without a consensus are represented by a hyphen. The burial residue index was generated by SAM_T08 program.*

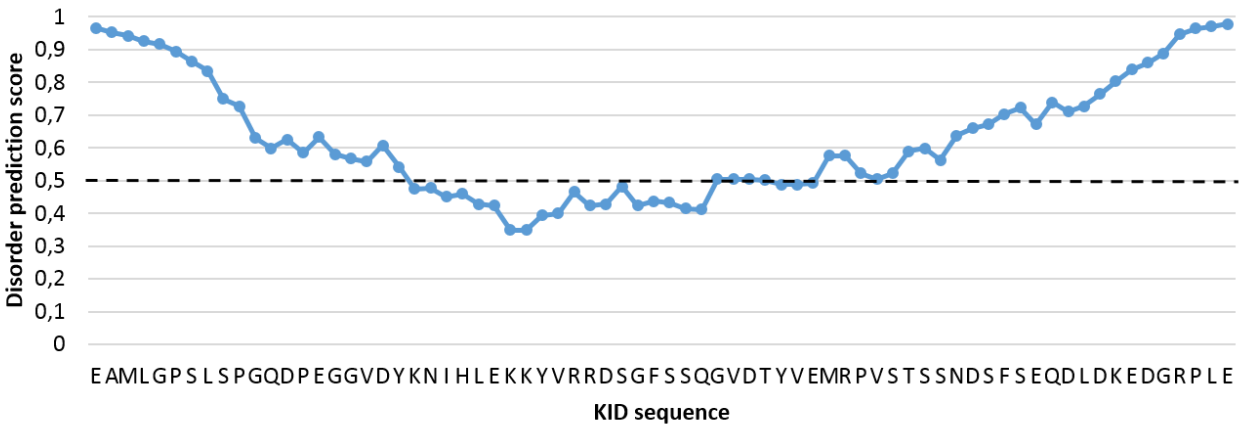


Figure A 5: ***Prediction of the disorder tendency for the CSF-1R's KID residues calculated by the IUPRED web-server.*** *Values above 0.5 indicate disordered structures* (DOSZTÁNYI *et al.*, 2005).

In possession of the bioinformatics analysis for the KID, we decided to model the CSF-1R^{full} by using <u>two</u> different protocols. Protocol 1 was reproduced from the one used by Isaure de Beauchêne during her PhD (CHAUVOT DE BEAUCHÊNE, 2013) for modeling the full-length KIT receptor. It consisted basically of using the *AbinitioRelax* module of Rosetta (LEAVER-FAY *et al.*, 2011) for the *de novo* prediction of the KID sequence (residues 680-751) before inserting it into the TK domain.

We have generated 10.000 models using Rosetta. Due to their very variable nature, clustering the models resulted in single structure groups, so we decided to simply sort the structures in function of their energy score. From the 100 lowest energy structures, models that did not correspond to a distance criteria of 13 Å between the N- and C-terminals were excluded. This selection was necessary to couple the model's extremities with the TK domain, avoiding clashes with the remaining protein residues. Six final models corresponded to the distance criteria (Fig. A 6).

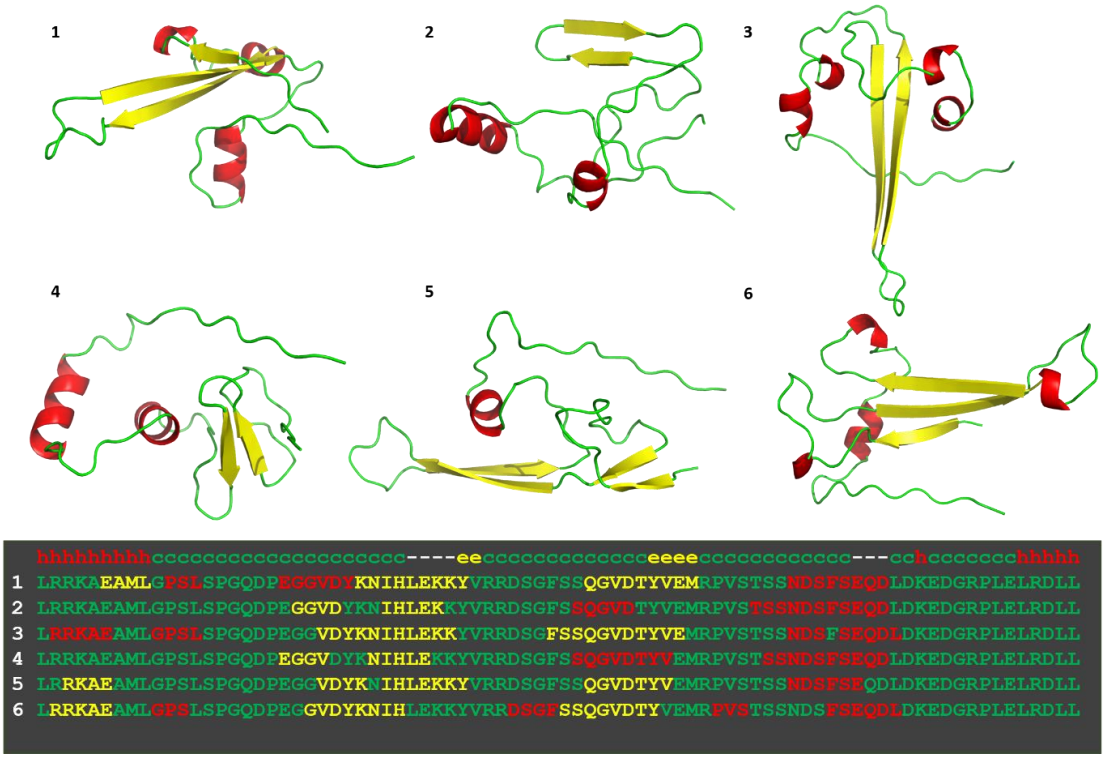


Figure A 6: ***Final models generated de novo by Rosetta****. The 10.000 models generated by the Abinitio.relax module were sorted by energy and the 100 lowest energy structures were analyzed using a distance criteria of 13 Å length between the N- and C- terminals. Only six models corresponded to the distance criteria. In this figure, the models are numerated accordingly to their energy score from the lowest to the highest energy model. Structures are colored by their secondary structure: α-helices in red, β-sheets in yellow and coil in green. For comparison, the consensus secondary structure prediction is represented using the same color code from the models.*

As we can see at Figure A 6, there is a consensus secondary structure preference for two well-formed β-strands and at least one helical region. Which is somehow coherent with the secondary structure prediction for the KID sequence. All six models were inserted into the TK domain of the crystal structure of auto-inhibited CSF-1R (PDB ID: 2OGV) and 100 models were generated for each insertion. The best models generated with Modeller (ESWAR *et al.*, 2008) can be seen at figure A 7. The distance criteria used in the models choice proved to be very effective since the insertion of the models did not result in any clashes with the remaining TK domain.

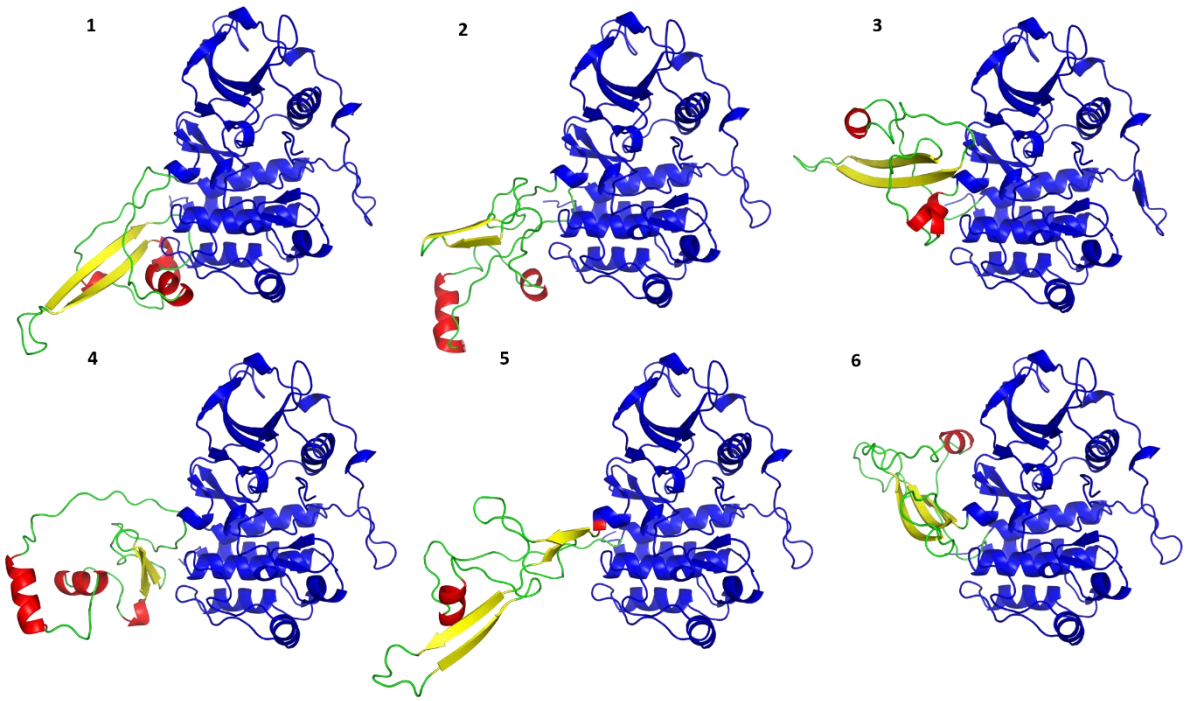


Figure A 7: ***Final models for the CSF-1Rf^{ull} generated using the Protocol 1***. *The models are numbered in the figure accordingly to figure A 6. The TK domain region modeled from the template 2OGV is represented in blue and the KID region predicted by Rosetta is colored by their secondary structure as in figure A 6.*

A recent study has compared different modeling techniques in order to stablish the best multi-method approach for the modeling of loops (JAMROZ & KOLINSKI, 2010). The authors have performed test predictions of loop regions of different sizes using MODELLER, ROSETTA and CABS (BLASZCZYK *et al.*, 2012), a coarse-grained *de novo* modeling tool.

The authors have also stablished a protocol in which they increased the model accuracy by combining Modeller and CABS, by using the 10 top ranked models from sets of 500 models generated by Modeller as multiple templates for CABS modeling.

In Protocol 2, we tried a similar approach by combining a previous step of comparative modeling, using Modeller followed by de novo prediction using two programs: Rosetta, using the *loopmodel* module; and CABS-fold (BLASZCZYK *et al.*, 2013). We used two templates for the comparative modeling step: the crystal structure of the auto-inhibited CSF-1R (PDB ID: 2OGV) and a crystal structure of inactive CSF-1R complexed with an inhibitor (PDB ID: 3LCO), retrieved earlier by BLAST, which contains a small solved portion of the KID.

Before constructing the model, we have manually altered the multiple templates-sequence alignment so that Modeller would use exclusively the 3LCO structure to construct the KID region. The best model, ranked by DOPE score was submitted to Rosetta and CABS-FOLD to reconstruct only the remaining KID region absent from the crystal 3LCO (residues 685-748), built as a solvent-exposed long loop by Modeller.

Due to the very long nature of Rosetta's calculations, only 431 models were generated in time for the finalization of this thesis. The models were clustered and the top seven structures, representatives of each cluster, are shown at Figure A 8. The KID is folded predominantly in coil. Differently from Rosetta, the models outputted by the CABS-fold are more structured and folded onto the TK domain (Fig. A 9). The folding is predominantly in α-helices with long beta sheets, similar to what we found in the first protocol using Rosetta.

The different nature of the *de novo* protocols used in this work had a consequence on the final constructed models. The *AbinitioRelax* module of Rosetta outputted fully folded structures but with no consensus among the results, besides the unviability of the application to model the full-length protein due to the associated computational cost. In addition, we cannot guarantee the validity of the approach applied in Protocol 1, since the modeling of the KID was done outside of the context of protein's TK domain.

The *de novo* reconstruction of the KID using the module *loopmodel*, also from Rosetta, has evidenced the inability of the program to solve structures with sequences bigger than a few amino acids (MANDELL; COUTSIAS & KORTEMME, 2009). CABS-fold web-server showed to be a viable option in terms of calculation time, since the server returns the results within a few hours, probably due to the coarse-grained representation of the protein before running the Monte Carlo Replica Exchange simulations. The use of a temperature range in the de *novo prediction* enhances the conformational sampling, which can improve the final models.

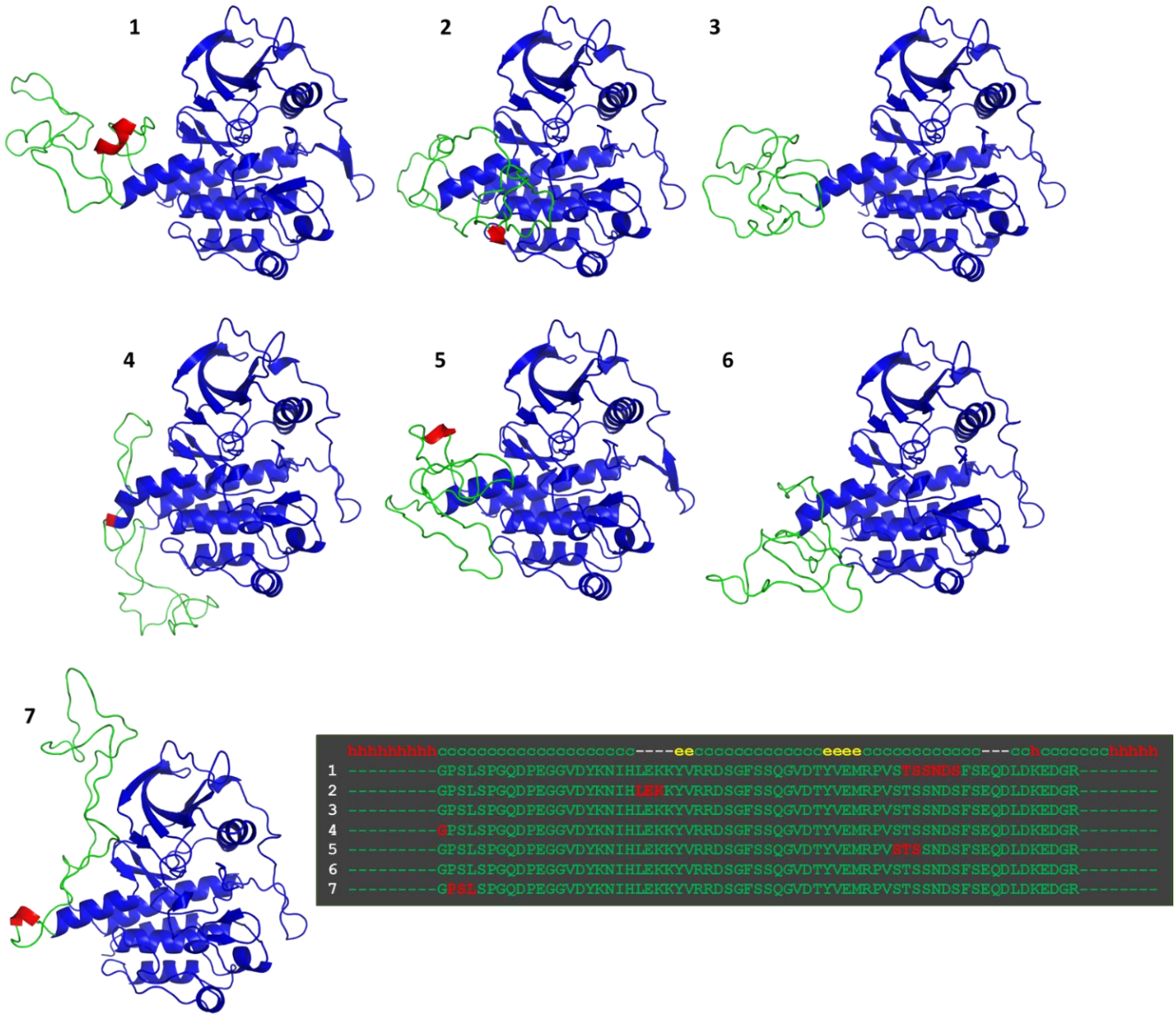
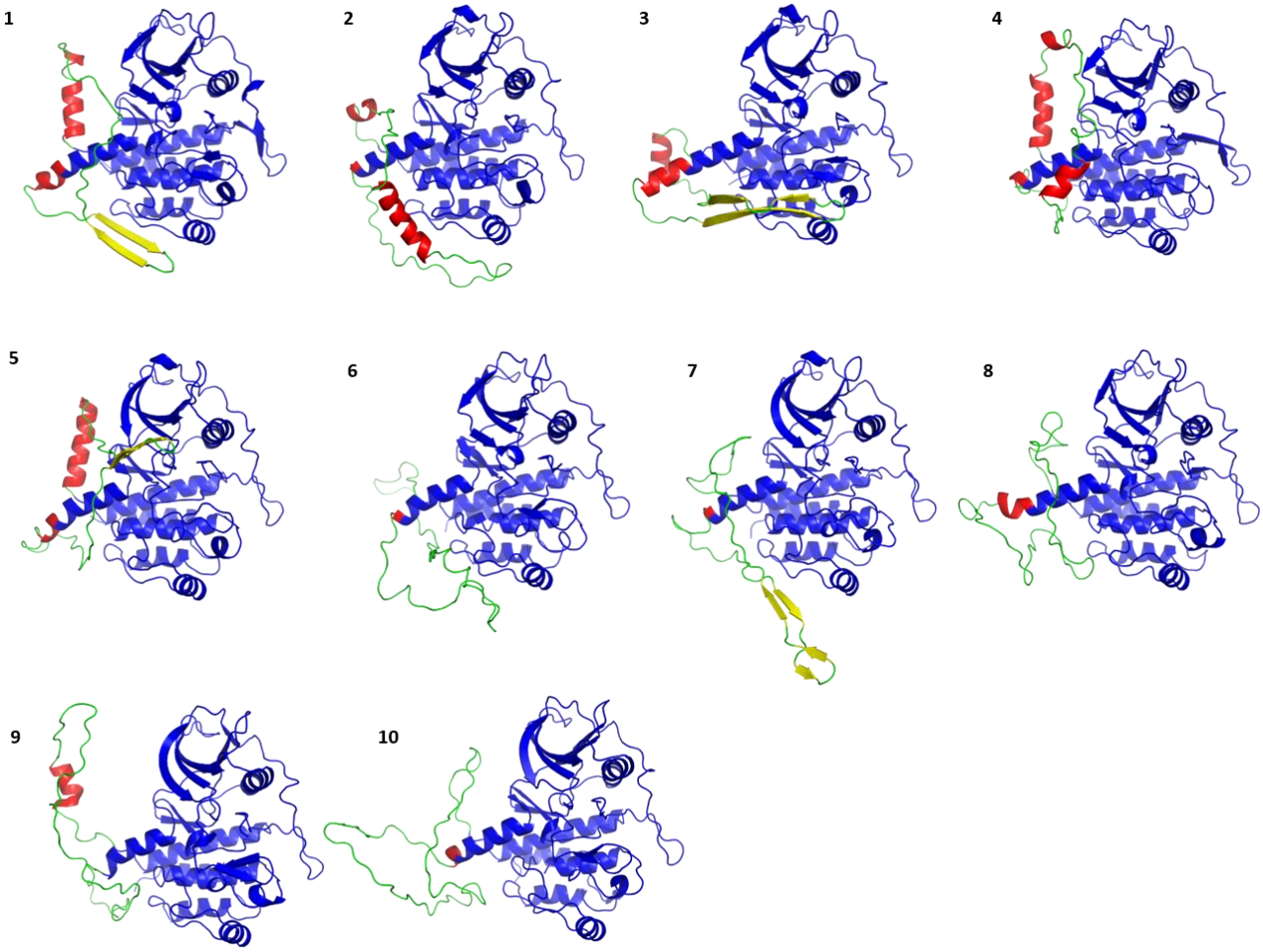


Figure A 8: ***Best models outputted by Rosetta following the Protocol 2****. The TK domain region modeled using Modeller is colored in blue and the KID predicted region is colored by their secondary structure as in figures 31 and 32. The regions denoted by – in the sequences are already modelled, using the PDB 3LCO as template.*




```
   hhhhhhhhhhcccccccccccccccccccc----eecccccccccccceeeeccccccccccc---cchccccccchhhhh
1  ---------GPSLSPGQDPEGGVDYKNIHLEKKYVRRDSGFSSQGVDTYVEMRPVSTSSNDSFSEQDLDKEDGR--------
2  ---------GPSLSPGQDPEGGVDYKNIHLEKKYVRRDSGFSSQGVDTYVEMRPVSTSSNDSFSEQDLDKEDGR--------
3  ---------GPSLSPGQDPEGGVDYKNIHLEKKYVRRDSGFSSQGVDTYVEMRPVSTSSNDSFSEQDLDKEDGR--------
4  ---------GPSLSPGQDPEGGVDYKNIHLEKKYVRRDSGFSSQGVDTYVEMRPVSTSSNDSFSEQDLDKEDGR--------
5  ---------GPSLSPGQDPEGGVDYKNIHLEKKYVRRDSGFSSQGVDTYVEMRPVSTSSNDSFSEQDLDKEDGR--------
6  ---------GPSLSPGQDPEGGVDYKNIHLEKKYVRRDSGFSSQGVDTYVEMRPVSTSSNDSFSEQDLDKEDGR--------
7  ---------GPSLSPGQDPEGGVDYKNIHLEKKYVRRDSGFSSQGVDTYVEMRPVSTSSNDSFSEQDLDKEDGR--------
8  ---------GPSLSPGQDPEGGVDYKNIHLEKKYVRRDSGFSSQGVDTYVEMRPVSTSSNDSFSEQDLDKEDGR--------
9  ---------GPSLSPGQDPEGGVDYKNIHLEKKYVRRDSGFSSQGVDTYVEMRPVSTSSNDSFSEQDLDKEDGR--------
10 ---------GPSLSPGQDPEGGVDYKNIHLEKKYVRRDSGFSSQGVDTYVEMRPVSTSSNDSFSEQDLDKEDGR--------
```

Figure A 9: ***Best models outputted by the CABS-fold web-server following the Protocol 2****. The TK domain region modeled using Modeller is colored in blue and the KID region predicted by CABS is colored by their secondary structure as in figures 31 and 32. The regions denoted by – in the sequences are already modelled, using the PDB 3LCO as template.*

Alternative modeling approaches, such as threading techniques (BOWIE; LUTHY & EISENBERG, 1991; JONES; TAYLOR & THORNTON, 1992), which use fold recognition could be applied to the problem. In addition, MD simulations should be applied to the best models of each approach in order to investigate the stability of the KID folding.

2. Simulations parameters file

Below, you'll find the main parameters file used in the docking, MD simulations and energy analysis. The topology and coordinate files of imatinib correspond to the parameters of the CHARMM27 forcefield. The charges were retrieved from a previous study (ALEKSANDROV & SIMONSON, 2010).

2.1. Docking

An example of input file used in the docking simulations performed at Maestro.

```
#  Multiple input structures can be specified by adding additional
#  INPUT_FILE lines or including multiple structures in a single
#  file.
#
#  If beginning with an existing Pose Viewer file, simply specify
#  it as the INPUT_FILE (making sure the name ends in "_pv.mae"
#  or "_pv.maegz") and ensure that the INITIAL_DOCKING stage
#  is commented out.  The ligand used in producing the Pose Viewer
#  file must also be provided to the GLIDE_DOCKING2 stage,
#  using the LIGAND_FILE keyword.

INPUT_FILE  FMS_D802V_nojmr.mae

# Prime Loop Prediction
#  Perform a loop prediction on the specified loop, including
#  side chains within the given distance.  Only return
#  structures within the specified energy range from the
#  lowest energy prediction, up to the maximum number of
#  conformations given.
#
#  Note: This stage is disabled by default.  Uncomment the
#   lines below and edit the fields appropriately to enable it.
#STAGE PRIME_LOOP
#  START_RESIDUE A:11
#  END_RESIDUE A:16
#  RES_SPHERE 7.5
#  MAX_ENERGY_GAP 30.0
#  MAX_STRUCTURES 5
#  USE_MEMBRANE no

STAGE VDW_SCALING
  BINDING_SITE ligand _:1

STAGE PREDICT_FLEXIBILITY
  BINDING_SITE ligand _:1

STAGE INITIAL_DOCKING
  BINDING_SITE ligand _:1
  INNERBOX 10.0
```

```
  OUTERBOX auto
  LIGAND_FILE imab.prot1.mae
  LIGANDS_TO_DOCK all
  VARIANTS_TO_RUN A,B,C,D,E,F,G
  DOCKING_RINGCONFCUT 2.5
  DOCKING_AMIDE_MODE penal

STAGE COMPILE_RESIDUE_LIST
  DISTANCE_CUTOFF 5.0

STAGE PRIME_REFINEMENT
  NUMBER_OF_PASSES  1
  USE_MEMBRANE no
  OPLS_VERSION OPLS_2005

STAGE GLIDE_DOCKING2
  BINDING_SITE ligand Z:999
  INNERBOX 5.0
  OUTERBOX auto
  LIGAND_FILE  imab.prot1.mae
  LIGANDS_TO_DOCK existing
  DOCKING_PRECISION SP
  DOCKING_RINGCONFCUT 2.5
  DOCKING_AMIDE_MODE penal

STAGE SCORING
  SCORE_NAME  r_psp_IFDScore
  TERM 1.000,r_psp_Prime_Energy,1
  TERM 9.057,r_i_glide_gscore,0
  TERM 1.428,r_i_glide_ecoul,0
  REPORT_FILE report.csv
```

2.2. MD simulations

Topology file for imatinib:

```
; Topology file
; Built itp for imatinib_FMS.mol2
;    by user pri     Wed May 21 11:20:35 CEST 2014
; ----
;

[ atomtypes ]
; name at.num  mass   charge  ptype    sigma            epsilon
CB      6   12.0110  0.0  A         0.355005    0.292880
NPYD    7   14.0067  0.0  A         0.329632    0.836800
CR      6   12.0110  0.0  A         0.387541    0.230120
NC=C    7   14.0067  0.0  A         0.329632    0.836800
NC=O    7   14.0067  0.0  A         0.329632    0.836800
C=O     6   12.0110  0.0  A         0.356359    0.460240
O=C     8   15.9994  0.0  A         0.302905    0.502080
NR      7   14.0067  0.0  A         0.329632    0.836800
NRP     7   14.0067  0.0  A         0.329632    0.836800
HCMM    1    1.0079  0.0  A         0.235197    0.092048
HNCO    1    1.0079  0.0  A         0.040001    0.192464
HNRP    1    1.0079  0.0  A         0.040001    0.192464
```

```
[ pairtypes ]
;  i     j    func     sigma1-4       epsilon1-4 ; THESE ARE 1-4
INTERACTIONS
CR       CB     1      0.346773    0.110698
CR       NPYD   1      0.334087    0.187114
CR       CR     1      0.338541    0.041840
CR       NC=C   1      0.334087    0.187114
CR       NC=O   1      0.334087    0.187114
CR       C=O    1      0.347450    0.138768
CR       O=C    1      0.293997    0.144938
CR       NR     1      0.334087    0.187114
CR       NRP    1      0.334087    0.187114
CR       HCMM   1      0.286869    0.062059
CR       HNCO   1      0.189271    0.089737
CR       HNRP   1      0.189271    0.089737
O=C      CB     1      0.302228    0.383470
O=C      NPYD   1      0.289542    0.648182
O=C      NC=C   1      0.289542    0.648182
O=C      NC=O   1      0.289542    0.648182
O=C      C=O    1      0.302905    0.480705
O=C      O=C    1      0.249452    0.502080
O=C      NR     1      0.289542    0.648182
O=C      NRP    1      0.289542    0.648182
O=C      HCMM   1      0.242324    0.214978
O=C      HNCO   1      0.144726    0.310857
O=C      HNRP   1      0.144726    0.310857

[ moleculetype ]
; Name nrexcl
imatinib_FMS 3

[ atoms ]
; nr type resnr resid atom cgnr charge mass
   1 CB   1  LIG C1      1  0.000   12.0110
   2 CB   1  LIG C2      2 -0.162   12.0110
   3 CB   1  LIG C3      3 -0.264   12.0110
   4 CB   1  LIG C4      4  0.291   12.0110
   5 CB   1  LIG C5      5 -0.264   12.0110
   6 CB   1  LIG C6      6  0.294   12.0110
   7 CB   1  LIG C7      7  0.984   12.0110
   8 NPYD 1  LIG N1      8 -0.738   14.0067
   9 CB   1  LIG C8      9  0.161   12.0110
  10 CB   1  LIG C9     10 -0.079   12.0110
  11 CB   1  LIG C10    11  0.321   12.0110
  12 NPYD 1  LIG N2     12 -0.738   14.0067
  13 CB   1  LIG C11    13  0.000   12.0110
  14 CB   1  LIG C12    14 -0.203   12.0110
  15 CB   1  LIG C13    15 -0.156   12.0110
  16 CB   1  LIG C14    16  0.154   12.0110
  17 NPYD 1  LIG N3     17 -0.626   14.0067
  18 CB   1  LIG C15    18  0.154   12.0110
  19 CB   1  LIG C16    19 -0.040   12.0110
  20 CB   1  LIG C17    20 -0.115   12.0110
  21 CB   1  LIG C18    21 -0.115   12.0110
  22 CB   1  LIG C19    22  0.000   12.0110
  23 CB   1  LIG C20    23 -0.115   12.0110
  24 CB   1  LIG C21    24 -0.115   12.0110
  25 CR   1  LIG C22    25 -0.270   12.0110
  26 NC=C 1  LIG N4     26 -0.845   14.0067
  27 NC=O 1  LIG N5     27 -0.801   14.0067
```

```
  28 C=O  1  LIG C23    28  0.862   12.0110
  29 O=C  1  LIG O1     29 -0.553   15.9994
  30 CR   1  LIG C24    30 -0.136   12.0110
  31 NR   1  LIG N6     31 -0.550   14.0067
  32 CR   1  LIG C25    32 -0.161   12.0110
  33 CR   1  LIG C26    33 -0.167   12.0110
  34 NRP  1  LIG N7     34 -0.668   14.0067
  35 CR   1  LIG C27    35 -0.167   12.0110
  36 CR   1  LIG C28    36 -0.161   12.0110
  37 CR   1  LIG C29    37 -0.357   12.0110
  38 HCMM 1  LIG H1     38  0.162   1.0079
  39 HCMM 1  LIG H2     39  0.201   1.0079
  40 HCMM 1  LIG H3     40  0.201   1.0079
  41 HCMM 1  LIG H4     41  0.160   1.0079
  42 HCMM 1  LIG H5     42  0.079   1.0079
  43 HCMM 1  LIG H6     43  0.203   1.0079
  44 HCMM 1  LIG H7     44  0.156   1.0079
  45 HCMM 1  LIG H8     45  0.159   1.0079
  46 HCMM 1  LIG H9     46  0.159   1.0079
  47 HCMM 1  LIG H10    47  0.115   1.0079
  48 HCMM 1  LIG H11    48  0.115   1.0079
  49 HCMM 1  LIG H12    49  0.115   1.0079
  50 HCMM 1  LIG H13    50  0.115   1.0079
  51 HCMM 1  LIG H14    51  0.090   1.0079
  52 HCMM 1  LIG H15    52  0.090   1.0079
  53 HCMM 1  LIG H16    53  0.090   1.0079
  54 HNCO 1  LIG H17    54  0.401   1.0079
  55 HNCO 1  LIG H18    55  0.367   1.0079
  56 HCMM 1  LIG H19    56  0.184   1.0079
  57 HCMM 1  LIG H20    57  0.184   1.0079
  58 HCMM 1  LIG H21    58  0.187   1.0079
  59 HCMM 1  LIG H22    59  0.187   1.0079
  60 HCMM 1  LIG H23    60  0.264   1.0079
  61 HCMM 1  LIG H24    61  0.264   1.0079
  62 HNRP 1  LIG H25    62  0.490   1.0079
  63 HCMM 1  LIG H26    63  0.264   1.0079
  64 HCMM 1  LIG H27    64  0.264   1.0079
  65 HCMM 1  LIG H28    65  0.187   1.0079
  66 HCMM 1  LIG H29    66  0.187   1.0079
  67 HCMM 1  LIG H30    67  0.235   1.0079
  68 HCMM 1  LIG H31    68  0.235   1.0079
  69 HCMM 1  LIG H32    69  0.235   1.0079

[ bonds ]
; ai aj fu b0 kb, b0 kb
 68  37 1 0.10930  287014.9  0.10930  287014.9
 61  33 1 0.10930  287014.9  0.10930  287014.9
 67  37 1 0.10930  287014.9  0.10930  287014.9
 37  69 1 0.10930  287014.9  0.10930  287014.9
 37  34 1 0.14800  231490.7  0.14800  231490.7
 62  34 1 0.10280  371144.2  0.10280  371144.2
 34  33 1 0.14800  231490.7  0.14800  231490.7
 34  35 1 0.14800  231490.7  0.14800  231490.7
 33  60 1 0.10930  287014.9  0.10930  287014.9
 33  32 1 0.15080  256422.3  0.15080  256422.3
 59  32 1 0.10930  287014.9  0.10930  287014.9
 58  32 1 0.10930  287014.9  0.10930  287014.9
 32  31 1 0.14510  306165.0  0.14510  306165.0
 64  35 1 0.10930  287014.9  0.10930  287014.9
 35  36 1 0.15080  256422.3  0.15080  256422.3
 35  63 1 0.10930  287014.9  0.10930  287014.9
```

```
 65  36 1 0.10930  287014.9  0.10930  287014.9
 36  31 1 0.14510  306165.0  0.14510  306165.0
 36  66 1 0.10930  287014.9  0.10930  287014.9
 31  30 1 0.14510  306165.0  0.14510  306165.0
 57  30 1 0.10930  287014.9  0.10930  287014.9
 56  30 1 0.10930  287014.9  0.10930  287014.9
 30  22 1 0.14860  298517.5  0.14860  298517.5
 48  21 1 0.10840  319534.6  0.10840  319534.6
 22  21 1 0.13740  335613.7  0.13740  335613.7
 22  23 1 0.13740  335613.7  0.13740  335613.7
 21  20 1 0.13740  335613.7  0.13740  335613.7
 49  23 1 0.10840  319534.6  0.10840  319534.6
 23  24 1 0.13740  335613.7  0.13740  335613.7
 20  47 1 0.10840  319534.6  0.10840  319534.6
 20  19 1 0.13740  335613.7  0.13740  335613.7
 24  19 1 0.13740  335613.7  0.13740  335613.7
 24  50 1 0.10840  319534.6  0.10840  319534.6
 19  28 1 0.14570  270273.8  0.14570  270273.8
 28  29 1 0.12220  779866.6  0.12220  779866.6
 28  27 1 0.13690  351030.1  0.13690  351030.1
 55  27 1 0.10150  401254.8  0.10150  401254.8
 27   4 1 0.13950  330133.5  0.13950  330133.5
  4   3 1 0.13740  335613.7  0.13740  335613.7
  4   5 1 0.13740  335613.7  0.13740  335613.7
 39   3 1 0.10840  319534.6  0.10840  319534.6
 40   5 1 0.10840  319534.6  0.10840  319534.6
 41   9 1 0.10840  319534.6  0.10840  319534.6
  3   2 1 0.13740  335613.7  0.13740  335613.7
  5   6 1 0.13740  335613.7  0.13740  335613.7
  9   8 1 0.13330  345489.6  0.13330  345489.6
  9  10 1 0.13740  335613.7  0.13740  335613.7
  8   7 1 0.13330  345489.6  0.13330  345489.6
  2  38 1 0.10840  319534.6  0.10840  319534.6
  2   1 1 0.13740  335613.7  0.13740  335613.7
 42  10 1 0.10840  319534.6  0.10840  319534.6
  6   1 1 0.13740  335613.7  0.13740  335613.7
  6  26 1 0.13980  371445.5  0.13980  371445.5
 10  11 1 0.13740  335613.7  0.13740  335613.7
  7  26 1 0.13980  371445.5  0.13980  371445.5
  7  12 1 0.13330  345489.6  0.13330  345489.6
  1  25 1 0.14860  298517.5  0.14860  298517.5
 26  54 1 0.10180  396015.6  0.10180  396015.6
 11  12 1 0.13330  345489.6  0.13330  345489.6
 11  13 1 0.13740  335613.7  0.13740  335613.7
 43  14 1 0.10840  319534.6  0.10840  319534.6
 13  14 1 0.13740  335613.7  0.13740  335613.7
 13  18 1 0.13740  335613.7  0.13740  335613.7
 25  53 1 0.10930  287014.9  0.10930  287014.9
 25  51 1 0.10930  287014.9  0.10930  287014.9
 25  52 1 0.10930  287014.9  0.10930  287014.9
 14  15 1 0.13740  335613.7  0.13740  335613.7
 46  18 1 0.10840  319534.6  0.10840  319534.6
 18  17 1 0.13330  345489.6  0.13330  345489.6
 15  44 1 0.10840  319534.6  0.10840  319534.6
 15  16 1 0.13740  335613.7  0.13740  335613.7
 17  16 1 0.13330  345489.6  0.13330  345489.6
 16  45 1 0.10840  319534.6  0.10840  319534.6

[ pairs ]
; ai aj fu
  1   4 1
```

```
 1  39 1
 1  40 1
 1   7 1
 1  54 1
 2   5 1
 2  26 1
 2  51 1
 2  52 1
 2  53 1
 2  27 1
 3   6 1
 3  25 1
 3  40 1
 3  28 1
 3  55 1
 4  38 1
 4  26 1
 4  19 1
 4  29 1
 5  39 1
 5  28 1
 5  55 1
 5  25 1
 5   7 1
 5  54 1
 6  38 1
 6  51 1
 6  52 1
 6  53 1
 6  27 1
 6   8 1
 6  12 1
 7  10 1
 7  41 1
 7  13 1
 8  11 1
 8  54 1
 8  42 1
 9  12 1
 9  26 1
 9  13 1
10  14 1
10  18 1
11  41 1
11  26 1
11  15 1
11  43 1
11  17 1
11  46 1
12  54 1
12  42 1
12  14 1
12  18 1
13  42 1
13  16 1
13  44 1
14  17 1
14  46 1
14  45 1
15  18 1
16  43 1
```

```
16  46 1
17  44 1
18  43 1
18  45 1
19  22 1
19  48 1
19  49 1
19  55 1
20  23 1
20  50 1
20  27 1
20  29 1
20  30 1
21  24 1
21  28 1
21  49 1
21  31 1
21  56 1
21  57 1
22  47 1
22  50 1
22  32 1
22  36 1
23  48 1
23  31 1
23  56 1
23  57 1
23  28 1
24  47 1
24  27 1
24  29 1
24  30 1
25  38 1
25  26 1
26  40 1
27  39 1
27  40 1
28  47 1
28  50 1
29  55 1
30  48 1
30  49 1
30  33 1
30  58 1
30  59 1
30  35 1
30  65 1
30  66 1
31  34 1
31  60 1
31  61 1
31  63 1
31  64 1
32  56 1
32  57 1
32  35 1
32  65 1
32  66 1
32  37 1
32  62 1
33  36 1
```

```
 33   63 1
 33   64 1
 33   67 1
 33   68 1
 33   69 1
 34   58 1
 34   59 1
 34   65 1
 34   66 1
 35   60 1
 35   61 1
 35   67 1
 35   68 1
 35   69 1
 36   56 1
 36   57 1
 36   58 1
 36   59 1
 36   37 1
 36   62 1
 37   60 1
 37   61 1
 37   63 1
 37   64 1
 38   39 1
 41   42 1
 43   44 1
 44   45 1
 47   48 1
 49   50 1
 58   60 1
 58   61 1
 59   60 1
 59   61 1
 60   62 1
 61   62 1
 62   63 1
 62   64 1
 62   67 1
 62   68 1
 62   69 1
 63   65 1
 63   66 1
 64   65 1
 64   66 1

[ angles ]
; ai aj ak fu th0 kth ub0 kub th0 kth ub0 kub
  2   1   6 1  119.9770  402.88    119.9770  402.88
  2   1  25 1  120.4190  483.57    120.4190  483.57
  6   1  25 1  120.4190  483.57    120.4190  483.57
  1   2   3 1  119.9770  402.88    119.9770  402.88
  1   2  38 1  120.5710  339.05    120.5710  339.05
  3   2  38 1  120.5710  339.05    120.5710  339.05
  2   3   4 1  119.9770  402.88    119.9770  402.88
  2   3  39 1  120.5710  339.05    120.5710  339.05
  4   3  39 1  120.5710  339.05    120.5710  339.05
  3   4   5 1  119.9770  402.88    119.9770  402.88
  3   4  27 1  117.9180  617.27    117.9180  617.27
  5   4  27 1  117.9180  617.27    117.9180  617.27
  4   5   6 1  119.9770  402.88    119.9770  402.88
```

4	5	40	1	120.5710	339.05	120.5710	339.05
6	5	40	1	120.5710	339.05	120.5710	339.05
1	6	5	1	119.9770	402.88	119.9770	402.88
1	6	26	1	121.6330	629.31	121.6330	629.31
5	6	26	1	121.6330	629.31	121.6330	629.31
8	7	12	1	128.9380	436.60	128.9380	436.60
8	7	26	1	123.7550	616.66	123.7550	616.66
12	7	26	1	123.7550	616.66	123.7550	616.66
7	8	9	1	115.4060	653.40	115.4060	653.40
8	9	10	1	126.1390	358.92	126.1390	358.92
8	9	41	1	115.5880	417.33	115.5880	417.33
10	9	41	1	120.5710	339.05	120.5710	339.05
9	10	11	1	119.9770	402.88	119.9770	402.88
9	10	42	1	120.5710	339.05	120.5710	339.05
11	10	42	1	120.5710	339.05	120.5710	339.05
10	11	12	1	126.1390	358.92	126.1390	358.92
10	11	13	1	119.9770	402.88	119.9770	402.88
12	11	13	1	126.1390	358.92	126.1390	358.92
7	12	11	1	115.4060	653.40	115.4060	653.40
11	13	14	1	119.9770	402.88	119.9770	402.88
11	13	18	1	119.9770	402.88	119.9770	402.88
14	13	18	1	119.9770	402.88	119.9770	402.88
13	14	15	1	119.9770	402.88	119.9770	402.88
13	14	43	1	120.5710	339.05	120.5710	339.05
15	14	43	1	120.5710	339.05	120.5710	339.05
14	15	16	1	119.9770	402.88	119.9770	402.88
14	15	44	1	120.5710	339.05	120.5710	339.05
16	15	44	1	120.5710	339.05	120.5710	339.05
15	16	17	1	126.1390	358.92	126.1390	358.92
15	16	45	1	120.5710	339.05	120.5710	339.05
17	16	45	1	115.5880	417.33	115.5880	417.33
16	17	18	1	115.4060	653.40	115.4060	653.40
13	18	17	1	126.1390	358.92	126.1390	358.92
13	18	46	1	120.5710	339.05	120.5710	339.05
17	18	46	1	115.5880	417.33	115.5880	417.33
20	19	24	1	119.9770	402.88	119.9770	402.88
20	19	28	1	114.4750	480.57	114.4750	480.57
24	19	28	1	114.4750	480.57	114.4750	480.57
19	20	21	1	119.9770	402.88	119.9770	402.88
19	20	47	1	120.5710	339.05	120.5710	339.05
21	20	47	1	120.5710	339.05	120.5710	339.05
20	21	22	1	119.9770	402.88	119.9770	402.88
20	21	48	1	120.5710	339.05	120.5710	339.05
22	21	48	1	120.5710	339.05	120.5710	339.05
21	22	23	1	119.9770	402.88	119.9770	402.88
21	22	30	1	120.4190	483.57	120.4190	483.57
23	22	30	1	120.4190	483.57	120.4190	483.57
22	23	24	1	119.9770	402.88	119.9770	402.88
22	23	49	1	120.5710	339.05	120.5710	339.05
24	23	49	1	120.5710	339.05	120.5710	339.05
19	24	23	1	119.9770	402.88	119.9770	402.88
19	24	50	1	120.5710	339.05	120.5710	339.05
23	24	50	1	120.5710	339.05	120.5710	339.05
1	25	51	1	109.4910	377.58	109.4910	377.58
1	25	52	1	109.4910	377.58	109.4910	377.58
1	25	53	1	109.4910	377.58	109.4910	377.58
51	25	52	1	108.8360	310.74	108.8360	310.74
51	25	53	1	108.8360	310.74	108.8360	310.74
52	25	53	1	108.8360	310.74	108.8360	310.74
6	26	7	1	119.0180	604.62	119.0180	604.62
6	26	54	1	110.2880	398.66	110.2880	398.66

```
  7  26  54 1  110.2880  398.66    110.2880  398.66
  4  27  28 1  118.5960  616.06    118.5960  616.06
  4  27  55 1  118.2270  378.18    118.2270  378.18
 28  27  55 1  120.2770  346.27    120.2770  346.27
 19  28  27 1  112.4950  663.03    112.4950  663.03
 19  28  29 1  119.9680  442.02    119.9680  442.02
 27  28  29 1  127.1520  546.20    127.1520  546.20
 22  30  31 1  110.9920  656.41    110.9920  656.41
 22  30  56 1  109.4910  377.58    109.4910  377.58
 22  30  57 1  109.4910  377.58    109.4910  377.58
 31  30  56 1  110.2970  393.25    110.2970  393.25
 31  30  57 1  110.2970  393.25    110.2970  393.25
 56  30  57 1  108.8360  310.74    108.8360  310.74
 30  31  32 1  107.0180  656.41    107.0180  656.41
 30  31  36 1  107.0180  656.41    107.0180  656.41
 32  31  36 1  107.0180  656.41    107.0180  656.41
 31  32  33 1  108.2900  467.91    108.2900  467.91
 31  32  58 1  110.2970  393.25    110.2970  393.25
 31  32  59 1  110.2970  393.25    110.2970  393.25
 33  32  58 1  110.5490  383.00    110.5490  383.00
 33  32  59 1  110.5490  383.00    110.5490  383.00
 58  32  59 1  108.8360  310.74    108.8360  310.74
 32  33  34 1  106.4930  710.01    106.4930  710.01
 32  33  60 1  110.5490  383.00    110.5490  383.00
 32  33  61 1  110.5490  383.00    110.5490  383.00
 34  33  60 1  106.2240  525.13    106.2240  525.13
 34  33  61 1  106.2240  525.13    106.2240  525.13
 60  33  61 1  108.8360  310.74    108.8360  310.74
 33  34  35 1  112.2510  519.10    112.2510  519.10
 33  34  37 1  112.2510  519.10    112.2510  519.10
 33  34  62 1  111.2060  346.87    111.2060  346.87
 35  34  37 1  112.2510  519.10    112.2510  519.10
 35  34  62 1  111.2060  346.87    111.2060  346.87
 37  34  62 1  111.2060  346.87    111.2060  346.87
 34  35  36 1  106.4930  710.01    106.4930  710.01
 34  35  63 1  106.2240  525.13    106.2240  525.13
 34  35  64 1  106.2240  525.13    106.2240  525.13
 36  35  63 1  110.5490  383.00    110.5490  383.00
 36  35  64 1  110.5490  383.00    110.5490  383.00
 63  35  64 1  108.8360  310.74    108.8360  310.74
 31  36  35 1  108.2900  467.91    108.2900  467.91
 31  36  65 1  110.2970  393.25    110.2970  393.25
 31  36  66 1  110.2970  393.25    110.2970  393.25
 35  36  65 1  110.5490  383.00    110.5490  383.00
 35  36  66 1  110.5490  383.00    110.5490  383.00
 65  36  66 1  108.8360  310.74    108.8360  310.74
 34  37  67 1  106.2240  525.13    106.2240  525.13
 34  37  68 1  106.2240  525.13    106.2240  525.13
 34  37  69 1  106.2240  525.13    106.2240  525.13
 67  37  68 1  108.8360  310.74    108.8360  310.74
 67  37  69 1  108.8360  310.74    108.8360  310.74
 68  37  69 1  108.8360  310.74    108.8360  310.74

[ dihedrals ]
; ai aj ak al fu phi0 kphi mult phi0 kphi mult
  1   2   3   4 9 180.00  14.6440 2   180.00  14.6440 2
  1   2   3  39 9 180.00  14.6440 2   180.00  14.6440 2
  1   6   5   4 9 180.00  14.6440 2   180.00  14.6440 2
  1   6   5  40 9 180.00  14.6440 2   180.00  14.6440 2
  1   6  26   7 9 180.00   8.3680 2   180.00   8.3680 2
  1   6  26  54 9   0.00   1.4937 1     0.00   1.4937 1
```

```
 1   6  26  54 9 180.00   5.4978 2   180.00   5.4978 2
 1   6  26  54 9   0.00   7.0166 3     0.00   7.0166 3
 2   1   6   5 9 180.00  14.6440 2   180.00  14.6440 2
 2   1   6  26 9 180.00  14.6440 2   180.00  14.6440 2
 2   1  25  51 9 180.00  -0.8786 2   180.00  -0.8786 2
 2   1  25  51 9   0.00   0.8201 3     0.00   0.8201 3
 2   1  25  52 9 180.00  -0.8786 2   180.00  -0.8786 2
 2   1  25  52 9   0.00   0.8201 3     0.00   0.8201 3
 2   1  25  53 9 180.00  -0.8786 2   180.00  -0.8786 2
 2   1  25  53 9   0.00   0.8201 3     0.00   0.8201 3
 2   3   4   5 9 180.00  14.6440 2   180.00  14.6440 2
 2   3   4  27 9 180.00  14.6440 2   180.00  14.6440 2
 3   2   1   6 9 180.00  14.6440 2   180.00  14.6440 2
 3   2   1  25 9 180.00  14.6440 2   180.00  14.6440 2
 3   4   5   6 9 180.00  14.6440 2   180.00  14.6440 2
 3   4   5  40 9 180.00  14.6440 2   180.00  14.6440 2
 3   4  27  28 9 180.00  12.5520 2   180.00  12.5520 2
 3   4  27  55 9 180.00  12.5520 2   180.00  12.5520 2
 4   3   2  38 9 180.00  14.6440 2   180.00  14.6440 2
 4   5   6  26 9 180.00  14.6440 2   180.00  14.6440 2
 4  27  28  19 9 180.00  12.5520 2   180.00  12.5520 2
 4  27  28  29 9 180.00  12.5520 2   180.00  12.5520 2
 5   4   3  39 9 180.00  14.6440 2   180.00  14.6440 2
 5   4  27  28 9 180.00  12.5520 2   180.00  12.5520 2
 5   4  27  55 9 180.00  12.5520 2   180.00  12.5520 2
 5   6   1  25 9 180.00  14.6440 2   180.00  14.6440 2
 5   6  26   7 9 180.00   8.3680 2   180.00   8.3680 2
 5   6  26  54 9   0.00   1.4937 1     0.00   1.4937 1
 5   6  26  54 9 180.00   5.4978 2   180.00   5.4978 2
 5   6  26  54 9   0.00   7.0166 3     0.00   7.0166 3
 6   1   2  38 9 180.00  14.6440 2   180.00  14.6440 2
 6   1  25  51 9 180.00  -0.8786 2   180.00  -0.8786 2
 6   1  25  51 9   0.00   0.8201 3     0.00   0.8201 3
 6   1  25  52 9 180.00  -0.8786 2   180.00  -0.8786 2
 6   1  25  52 9   0.00   0.8201 3     0.00   0.8201 3
 6   1  25  53 9 180.00  -0.8786 2   180.00  -0.8786 2
 6   1  25  53 9   0.00   0.8201 3     0.00   0.8201 3
 6   5   4  27 9 180.00  14.6440 2   180.00  14.6440 2
 6  26   7   8 9 180.00   8.3680 2   180.00   8.3680 2
 6  26   7  12 9 180.00   8.3680 2   180.00   8.3680 2
 7   8   9  10 9 180.00  14.6440 2   180.00  14.6440 2
 7   8   9  41 9 180.00  14.6440 2   180.00  14.6440 2
 7  12  11  10 9 180.00  14.6440 2   180.00  14.6440 2
 7  12  11  13 9 180.00  14.6440 2   180.00  14.6440 2
 8   7  12  11 9 180.00  14.6440 2   180.00  14.6440 2
 8   7  26  54 9 180.00   8.3680 2   180.00   8.3680 2
 8   9  10  11 9 180.00  14.6440 2   180.00  14.6440 2
 8   9  10  42 9 180.00  14.6440 2   180.00  14.6440 2
 9   8   7  12 9 180.00  14.6440 2   180.00  14.6440 2
 9   8   7  26 9 180.00  14.6440 2   180.00  14.6440 2
 9  10  11  12 9 180.00  14.6440 2   180.00  14.6440 2
 9  10  11  13 9 180.00  14.6440 2   180.00  14.6440 2
10  11  13  14 9 180.00  14.6440 2   180.00  14.6440 2
10  11  13  18 9 180.00  14.6440 2   180.00  14.6440 2
11  10   9  41 9 180.00  14.6440 2   180.00  14.6440 2
11  12   7  26 9 180.00  14.6440 2   180.00  14.6440 2
11  13  14  15 9 180.00  14.6440 2   180.00  14.6440 2
11  13  14  43 9 180.00  14.6440 2   180.00  14.6440 2
11  13  18  17 9 180.00  14.6440 2   180.00  14.6440 2
11  13  18  46 9 180.00  14.6440 2   180.00  14.6440 2
12   7  26  54 9 180.00   8.3680 2   180.00   8.3680 2
```

```
12  11  10  42 9 180.00  14.6440 2   180.00  14.6440 2
12  11  13  14 9 180.00  14.6440 2   180.00  14.6440 2
12  11  13  18 9 180.00  14.6440 2   180.00  14.6440 2
13  11  10  42 9 180.00  14.6440 2   180.00  14.6440 2
13  14  15  16 9 180.00  14.6440 2   180.00  14.6440 2
13  14  15  44 9 180.00  14.6440 2   180.00  14.6440 2
13  18  17  16 9 180.00  14.6440 2   180.00  14.6440 2
14  13  18  17 9 180.00  14.6440 2   180.00  14.6440 2
14  13  18  46 9 180.00  14.6440 2   180.00  14.6440 2
14  15  16  17 9 180.00  14.6440 2   180.00  14.6440 2
14  15  16  45 9 180.00  14.6440 2   180.00  14.6440 2
15  14  13  18 9 180.00  14.6440 2   180.00  14.6440 2
15  16  17  18 9 180.00  14.6440 2   180.00  14.6440 2
16  15  14  43 9 180.00  14.6440 2   180.00  14.6440 2
16  17  18  46 9 180.00  14.6440 2   180.00  14.6440 2
17  16  15  44 9 180.00  14.6440 2   180.00  14.6440 2
18  13  14  43 9 180.00  14.6440 2   180.00  14.6440 2
18  17  16  45 9 180.00  14.6440 2   180.00  14.6440 2
19  20  21  22 9 180.00  14.6440 2   180.00  14.6440 2
19  20  21  48 9 180.00  14.6440 2   180.00  14.6440 2
19  24  23  22 9 180.00  14.6440 2   180.00  14.6440 2
19  24  23  49 9 180.00  14.6440 2   180.00  14.6440 2
19  28  27  55 9 180.00  12.5520 2   180.00  12.5520 2
20  19  24  23 9 180.00  14.6440 2   180.00  14.6440 2
20  19  24  50 9 180.00  14.6440 2   180.00  14.6440 2
20  19  28  27 9 180.00   5.2300 2   180.00   5.2300 2
20  19  28  29 9 180.00   4.7196 2   180.00   4.7196 2
20  21  22  23 9 180.00  14.6440 2   180.00  14.6440 2
20  21  22  30 9 180.00  14.6440 2   180.00  14.6440 2
21  20  19  24 9 180.00  14.6440 2   180.00  14.6440 2
21  20  19  28 9 180.00  14.6440 2   180.00  14.6440 2
21  22  23  24 9 180.00  14.6440 2   180.00  14.6440 2
21  22  23  49 9 180.00  14.6440 2   180.00  14.6440 2
21  22  30  31 9   0.00   0.4184 3     0.00   0.4184 3
21  22  30  56 9 180.00  -0.8786 2   180.00  -0.8786 2
21  22  30  56 9   0.00   0.8201 3     0.00   0.8201 3
21  22  30  57 9 180.00  -0.8786 2   180.00  -0.8786 2
21  22  30  57 9   0.00   0.8201 3     0.00   0.8201 3
22  21  20  47 9 180.00  14.6440 2   180.00  14.6440 2
22  23  24  50 9 180.00  14.6440 2   180.00  14.6440 2
22  30  31  32 9 180.00  -0.6276 2   180.00  -0.6276 2
22  30  31  32 9   0.00   1.0460 3     0.00   1.0460 3
22  30  31  36 9 180.00  -0.6276 2   180.00  -0.6276 2
22  30  31  36 9   0.00   1.0460 3     0.00   1.0460 3
23  22  21  48 9 180.00  14.6440 2   180.00  14.6440 2
23  22  30  31 9   0.00   0.4184 3     0.00   0.4184 3
23  22  30  56 9 180.00  -0.8786 2   180.00  -0.8786 2
23  22  30  56 9   0.00   0.8201 3     0.00   0.8201 3
23  22  30  57 9 180.00  -0.8786 2   180.00  -0.8786 2
23  22  30  57 9   0.00   0.8201 3     0.00   0.8201 3
23  24  19  28 9 180.00  14.6440 2   180.00  14.6440 2
24  19  20  47 9 180.00  14.6440 2   180.00  14.6440 2
24  19  28  27 9 180.00   5.2300 2   180.00   5.2300 2
24  19  28  29 9 180.00   4.7196 2   180.00   4.7196 2
24  23  22  30 9 180.00  14.6440 2   180.00  14.6440 2
25   1   2  38 9 180.00  14.6440 2   180.00  14.6440 2
25   1   6  26 9 180.00  14.6440 2   180.00  14.6440 2
26   6   5  40 9 180.00  14.6440 2   180.00  14.6440 2
27   4   3  39 9 180.00  14.6440 2   180.00  14.6440 2
27   4   5  40 9 180.00  14.6440 2   180.00  14.6440 2
28  19  20  47 9 180.00   4.1840 2   180.00   4.1840 2
```

```
28  19  24  50 9 180.00   4.1840 2   180.00   4.1840 2
29  28  27  55 9   0.00   3.0041 1     0.00   3.0041 1
29  28  27  55 9 180.00  10.4056 2   180.00  10.4056 2
29  28  27  55 9   0.00  -0.9498 3     0.00  -0.9498 3
30  22  21  48 9 180.00  14.6440 2   180.00  14.6440 2
30  22  23  49 9 180.00  14.6440 2   180.00  14.6440 2
30  31  32  33 9   0.00  -0.9205 1     0.00  -0.9205 1
30  31  32  33 9 180.00   1.6443 2   180.00   1.6443 2
30  31  32  33 9   0.00   0.5690 3     0.00   0.5690 3
30  31  32  58 9   0.00   0.8242 1     0.00   0.8242 1
30  31  32  58 9 180.00  -0.8075 2   180.00  -0.8075 2
30  31  32  58 9   0.00   1.1757 3     0.00   1.1757 3
30  31  32  59 9   0.00   0.8242 1     0.00   0.8242 1
30  31  32  59 9 180.00  -0.8075 2   180.00  -0.8075 2
30  31  32  59 9   0.00   1.1757 3     0.00   1.1757 3
30  31  36  35 9   0.00  -0.9205 1     0.00  -0.9205 1
30  31  36  35 9 180.00   1.6443 2   180.00   1.6443 2
30  31  36  35 9   0.00   0.5690 3     0.00   0.5690 3
30  31  36  65 9   0.00   0.8242 1     0.00   0.8242 1
30  31  36  65 9 180.00  -0.8075 2   180.00  -0.8075 2
30  31  36  65 9   0.00   1.1757 3     0.00   1.1757 3
30  31  36  66 9   0.00   0.8242 1     0.00   0.8242 1
30  31  36  66 9 180.00  -0.8075 2   180.00  -0.8075 2
30  31  36  66 9   0.00   1.1757 3     0.00   1.1757 3
31  32  33  34 9   0.00   0.6276 3     0.00   0.6276 3
31  32  33  60 9   0.00  -1.5564 1     0.00  -1.5564 1
31  32  33  60 9 180.00  -2.5857 2   180.00  -2.5857 2
31  32  33  60 9   0.00   0.7071 3     0.00   0.7071 3
31  32  33  61 9   0.00  -1.5564 1     0.00  -1.5564 1
31  32  33  61 9 180.00  -2.5857 2   180.00  -2.5857 2
31  32  33  61 9   0.00   0.7071 3     0.00   0.7071 3
31  36  35  34 9   0.00   0.6276 3     0.00   0.6276 3
31  36  35  63 9   0.00  -1.5564 1     0.00  -1.5564 1
31  36  35  63 9 180.00  -2.5857 2   180.00  -2.5857 2
31  36  35  63 9   0.00   0.7071 3     0.00   0.7071 3
31  36  35  64 9   0.00  -1.5564 1     0.00  -1.5564 1
31  36  35  64 9 180.00  -2.5857 2   180.00  -2.5857 2
31  36  35  64 9   0.00   0.7071 3     0.00   0.7071 3
32  31  30  56 9   0.00   0.8242 1     0.00   0.8242 1
32  31  30  56 9 180.00  -0.8075 2   180.00  -0.8075 2
32  31  30  56 9   0.00   1.1757 3     0.00   1.1757 3
32  31  30  57 9   0.00   0.8242 1     0.00   0.8242 1
32  31  30  57 9 180.00  -0.8075 2   180.00  -0.8075 2
32  31  30  57 9   0.00   1.1757 3     0.00   1.1757 3
32  31  36  35 9   0.00  -0.9205 1     0.00  -0.9205 1
32  31  36  35 9 180.00   1.6443 2   180.00   1.6443 2
32  31  36  35 9   0.00   0.5690 3     0.00   0.5690 3
32  31  36  65 9   0.00   0.8242 1     0.00   0.8242 1
32  31  36  65 9 180.00  -0.8075 2   180.00  -0.8075 2
32  31  36  65 9   0.00   1.1757 3     0.00   1.1757 3
32  31  36  66 9   0.00   0.8242 1     0.00   0.8242 1
32  31  36  66 9 180.00  -0.8075 2   180.00  -0.8075 2
32  31  36  66 9   0.00   1.1757 3     0.00   1.1757 3
32  33  34  35 9   0.00   0.5230 3     0.00   0.5230 3
32  33  34  37 9   0.00   0.5230 3     0.00   0.5230 3
32  33  34  62 9   0.00   0.3891 3     0.00   0.3891 3
33  32  31  36 9   0.00  -0.9205 1     0.00  -0.9205 1
33  32  31  36 9 180.00   1.6443 2   180.00   1.6443 2
33  32  31  36 9   0.00   0.5690 3     0.00   0.5690 3
33  34  35  36 9   0.00   0.5230 3     0.00   0.5230 3
33  34  35  63 9   0.00   0.5146 3     0.00   0.5146 3
```

```
33  34  35  64 9   0.00   0.5146 3     0.00   0.5146 3
33  34  37  67 9   0.00   0.5146 3     0.00   0.5146 3
33  34  37  68 9   0.00   0.5146 3     0.00   0.5146 3
33  34  37  69 9   0.00   0.5146 3     0.00   0.5146 3
34  33  32  58 9   0.00   1.4477 1     0.00   1.4477 1
34  33  32  58 9 180.00  -1.1088 2   180.00  -1.1088 2
34  33  32  58 9   0.00   0.5816 3     0.00   0.5816 3
34  33  32  59 9   0.00   1.4477 1     0.00   1.4477 1
34  33  32  59 9 180.00  -1.1088 2   180.00  -1.1088 2
34  33  32  59 9   0.00   0.5816 3     0.00   0.5816 3
34  35  36  65 9   0.00   1.4477 1     0.00   1.4477 1
34  35  36  65 9 180.00  -1.1088 2   180.00  -1.1088 2
34  35  36  65 9   0.00   0.5816 3     0.00   0.5816 3
34  35  36  66 9   0.00   1.4477 1     0.00   1.4477 1
34  35  36  66 9 180.00  -1.1088 2   180.00  -1.1088 2
34  35  36  66 9   0.00   0.5816 3     0.00   0.5816 3
35  34  33  60 9   0.00   0.5146 3     0.00   0.5146 3
35  34  33  61 9   0.00   0.5146 3     0.00   0.5146 3
35  34  37  67 9   0.00   0.5146 3     0.00   0.5146 3
35  34  37  68 9   0.00   0.5146 3     0.00   0.5146 3
35  34  37  69 9   0.00   0.5146 3     0.00   0.5146 3
36  31  30  56 9   0.00   0.8242 1     0.00   0.8242 1
36  31  30  56 9 180.00  -0.8075 2   180.00  -0.8075 2
36  31  30  56 9   0.00   1.1757 3     0.00   1.1757 3
36  31  30  57 9   0.00   0.8242 1     0.00   0.8242 1
36  31  30  57 9 180.00  -0.8075 2   180.00  -0.8075 2
36  31  30  57 9   0.00   1.1757 3     0.00   1.1757 3
36  31  32  58 9   0.00   0.8242 1     0.00   0.8242 1
36  31  32  58 9 180.00  -0.8075 2   180.00  -0.8075 2
36  31  32  58 9   0.00   1.1757 3     0.00   1.1757 3
36  31  32  59 9   0.00   0.8242 1     0.00   0.8242 1
36  31  32  59 9 180.00  -0.8075 2   180.00  -0.8075 2
36  31  32  59 9   0.00   1.1757 3     0.00   1.1757 3
36  35  34  37 9   0.00   0.5230 3     0.00   0.5230 3
36  35  34  62 9   0.00   0.3891 3     0.00   0.3891 3
37  34  33  60 9   0.00   0.5146 3     0.00   0.5146 3
37  34  33  61 9   0.00   0.5146 3     0.00   0.5146 3
37  34  35  63 9   0.00   0.5146 3     0.00   0.5146 3
37  34  35  64 9   0.00   0.5146 3     0.00   0.5146 3
38   2   3  39 9 180.00  14.6440 2   180.00  14.6440 2
41   9  10  42 9 180.00  14.6440 2   180.00  14.6440 2
43  14  15  44 9 180.00  14.6440 2   180.00  14.6440 2
44  15  16  45 9 180.00  14.6440 2   180.00  14.6440 2
47  20  21  48 9 180.00  14.6440 2   180.00  14.6440 2
49  23  24  50 9 180.00  14.6440 2   180.00  14.6440 2
58  32  33  60 9   0.00   0.5941 1     0.00   0.5941 1
58  32  33  60 9 180.00  -2.8995 2   180.00  -2.8995 2
58  32  33  60 9   0.00   0.6569 3     0.00   0.6569 3
58  32  33  61 9   0.00   0.5941 1     0.00   0.5941 1
58  32  33  61 9 180.00  -2.8995 2   180.00  -2.8995 2
58  32  33  61 9   0.00   0.6569 3     0.00   0.6569 3
59  32  33  60 9   0.00   0.5941 1     0.00   0.5941 1
59  32  33  60 9 180.00  -2.8995 2   180.00  -2.8995 2
59  32  33  60 9   0.00   0.6569 3     0.00   0.6569 3
59  32  33  61 9   0.00   0.5941 1     0.00   0.5941 1
59  32  33  61 9 180.00  -2.8995 2   180.00  -2.8995 2
59  32  33  61 9   0.00   0.6569 3     0.00   0.6569 3
60  33  34  62 9   0.00   0.5439 3     0.00   0.5439 3
61  33  34  62 9   0.00   0.5439 3     0.00   0.5439 3
62  34  35  63 9   0.00   0.5439 3     0.00   0.5439 3
62  34  35  64 9   0.00   0.5439 3     0.00   0.5439 3
```

```
 62   34   37   67 9    0.00    0.5439 3      0.00    0.5439 3
 62   34   37   68 9    0.00    0.5439 3      0.00    0.5439 3
 62   34   37   69 9    0.00    0.5439 3      0.00    0.5439 3
 63   35   36   65 9    0.00    0.5941 1      0.00    0.5941 1
 63   35   36   65 9  180.00   -2.8995 2    180.00   -2.8995 2
 63   35   36   65 9    0.00    0.6569 3      0.00    0.6569 3
 63   35   36   66 9    0.00    0.5941 1      0.00    0.5941 1
 63   35   36   66 9  180.00   -2.8995 2    180.00   -2.8995 2
 63   35   36   66 9    0.00    0.6569 3      0.00    0.6569 3
 64   35   36   65 9    0.00    0.5941 1      0.00    0.5941 1
 64   35   36   65 9  180.00   -2.8995 2    180.00   -2.8995 2
 64   35   36   65 9    0.00    0.6569 3      0.00    0.6569 3
 64   35   36   66 9    0.00    0.5941 1      0.00    0.5941 1
 64   35   36   66 9  180.00   -2.8995 2    180.00   -2.8995 2
 64   35   36   66 9    0.00    0.6569 3      0.00    0.6569 3

[ dihedrals ]
; ai aj ak al fu xi0 kxi xi0 kxi
  1    2    6   25 2    0.00   24.0915      0.00   24.0915
  2    3    1   38 2    0.00    9.0291      0.00    9.0291
  3    4    2   39 2    0.00    9.0291      0.00    9.0291
  4   27    3    5 2    0.00   21.0790      0.00   21.0790
  6    5    1   26 2    0.00   27.6981      0.00   27.6981
 26    7    6   54 2    0.00   -3.0125      0.00   -3.0125
  7   12   26    8 2    0.00   21.0790      0.00   21.0790
  9   10    8   41 2    0.00   27.6981      0.00   27.6981
 10   11    9   42 2    0.00    9.0291      0.00    9.0291
 11   12   10   13 2    0.00   21.0790      0.00   21.0790
 13   18   11   14 2    0.00   21.0790      0.00   21.0790
 14   15   13   43 2    0.00    9.0291      0.00    9.0291
 15   16   14   44 2    0.00    9.0291      0.00    9.0291
 18   17   13   46 2    0.00   27.6981      0.00   27.6981
 27   28    4   55 2    0.00  -12.0416      0.00  -12.0416
 28   19   27   29 2    0.00   78.2910      0.00   78.2910
 19   24   28   20 2    0.00   16.2590      0.00   16.2590
 20   21   19   47 2    0.00    9.0291      0.00    9.0291
 21   22   20   48 2    0.00    9.0291      0.00    9.0291
 22   30   21   23 2    0.00   24.0915      0.00   24.0915
 30   31   22   57 2    0.00    0.0000      0.00    0.0000
 30   31   22   56 2    0.00    0.0000      0.00    0.0000
 31   36   30   32 2    0.00    0.0000      0.00    0.0000
 32   33   31   59 2    0.00    0.0000      0.00    0.0000
 32   33   31   58 2    0.00    0.0000      0.00    0.0000
 33   34   32   61 2    0.00    0.0000      0.00    0.0000
 33   61   32   60 2    0.00    0.0000      0.00    0.0000
 34   37   33   35 2    0.00    0.0000      0.00    0.0000
 34   37   33   62 2    0.00    0.0000      0.00    0.0000
  5    6    4   40 2    0.00    9.0291      0.00    9.0291
 16   17   15   45 2    0.00   27.6981      0.00   27.6981
 23   24   22   49 2    0.00    9.0291      0.00    9.0291
 24   23   19   50 2    0.00    9.0291      0.00    9.0291
 25   53    1   51 2    0.00    0.0000      0.00    0.0000
 25   53    1   52 2    0.00    0.0000      0.00    0.0000
 35   36   34   64 2    0.00    0.0000      0.00    0.0000
 35   64   34   63 2    0.00    0.0000      0.00    0.0000
 36   35   31   65 2    0.00    0.0000      0.00    0.0000
 36   35   31   66 2    0.00    0.0000      0.00    0.0000
 37   68   34   67 2    0.00    0.0000      0.00    0.0000
 37   68   34   69 2    0.00    0.0000      0.00    0.0000
```

```
#ifdef POSRES_LIGAND
[ position_restraints ]
; atom  type      fx      fy      fz
   1 1 1000 1000 1000
   2 1 1000 1000 1000
   3 1 1000 1000 1000
   4 1 1000 1000 1000
   5 1 1000 1000 1000
   6 1 1000 1000 1000
   7 1 1000 1000 1000
   8 1 1000 1000 1000
   9 1 1000 1000 1000
  10 1 1000 1000 1000
  11 1 1000 1000 1000
  12 1 1000 1000 1000
  13 1 1000 1000 1000
  14 1 1000 1000 1000
  15 1 1000 1000 1000
  16 1 1000 1000 1000
  17 1 1000 1000 1000
  18 1 1000 1000 1000
  19 1 1000 1000 1000
  20 1 1000 1000 1000
  21 1 1000 1000 1000
  22 1 1000 1000 1000
  23 1 1000 1000 1000
  24 1 1000 1000 1000
  25 1 1000 1000 1000
  26 1 1000 1000 1000
  27 1 1000 1000 1000
  28 1 1000 1000 1000
  29 1 1000 1000 1000
  30 1 1000 1000 1000
  31 1 1000 1000 1000
  32 1 1000 1000 1000
  33 1 1000 1000 1000
  34 1 1000 1000 1000
  35 1 1000 1000 1000
  36 1 1000 1000 1000
  37 1 1000 1000 1000
#endif
```

Coordinates file of imatinib:

```
TITLE     Gyas ROwers Mature At Cryogenic Speed
MODEL        1
HETATM    1  C1  UNK Z   1      24.023  26.589  44.587  1.00  0.00           C
HETATM    2  C2  UNK Z   1      22.837  26.589  43.827  1.00  0.00           C
HETATM    3  C3  UNK Z   1      22.872  26.367  42.442  1.00  0.00           C
HETATM    4  C4  UNK Z   1      24.090  26.104  41.795  1.00  0.00           C
HETATM    5  C5  UNK Z   1      25.290  26.133  42.549  1.00  0.00           C
HETATM    6  C6  UNK Z   1      25.262  26.369  43.941  1.00  0.00           C
HETATM    7  C7  UNK Z   1      27.746  26.655  44.480  1.00  0.00           C
HETATM    8  N1  UNK Z   1      28.026  27.465  43.463  1.00  0.00           N
HETATM    9  C8  UNK Z   1      29.323  27.790  43.292  1.00  0.00           C
HETATM   10  C9  UNK Z   1      30.330  27.287  44.137  1.00  0.00           C
HETATM   11  C10 UNK Z   1      29.923  26.421  45.176  1.00  0.00           C
HETATM   12  N2  UNK Z   1      28.612  26.136  45.344  1.00  0.00           N
HETATM   13  C11 UNK Z   1      30.863  25.750  46.102  1.00  0.00           C
HETATM   14  C12 UNK Z   1      32.177  26.225  46.313  1.00  0.00           C
HETATM   15  C13 UNK Z   1      33.045  25.528  47.175  1.00  0.00           C
```

```
HETATM   16  C14 UNK Z   1      32.570  24.367  47.811  1.00  0.00           C
HETATM   17  N3  UNK Z   1      31.322  23.891  47.631  1.00  0.00           N
HETATM   18  C15 UNK Z   1      30.501  24.573  46.803  1.00  0.00           C
HETATM   19  C16 UNK Z   1      24.503  25.298  38.124  1.00  0.00           C
HETATM   20  C17 UNK Z   1      25.182  24.382  37.290  1.00  0.00           C
HETATM   21  C18 UNK Z   1      24.783  24.205  35.952  1.00  0.00           C
HETATM   22  C19 UNK Z   1      23.705  24.952  35.433  1.00  0.00           C
HETATM   23  C20 UNK Z   1      23.015  25.857  36.265  1.00  0.00           C
HETATM   24  C21 UNK Z   1      23.414  26.035  37.601  1.00  0.00           C
HETATM   25  C22 UNK Z   1      23.935  26.813  46.089  1.00  0.00           C
HETATM   26  N4  UNK Z   1      26.438  26.289  44.696  1.00  0.00           N
HETATM   27  N5  UNK Z   1      24.004  25.801  40.411  1.00  0.00           N
HETATM   28  C23 UNK Z   1      24.977  25.518  39.529  1.00  0.00           C
HETATM   29  O1  UNK Z   1      26.170  25.389  39.803  1.00  0.00           O
HETATM   30  C24 UNK Z   1      23.304  24.807  33.979  1.00  0.00           C
HETATM   31  N6  UNK Z   1      23.364  26.066  33.237  1.00  0.00           N
HETATM   32  C25 UNK Z   1      22.829  25.840  31.898  1.00  0.00           C
HETATM   33  C26 UNK Z   1      22.809  27.165  31.107  1.00  0.00           C
HETATM   34  N7  UNK Z   1      24.170  27.750  31.011  1.00  0.00           N
HETATM   35  C27 UNK Z   1      24.776  27.879  32.358  1.00  0.00           C
HETATM   36  C28 UNK Z   1      24.750  26.528  33.099  1.00  0.00           C
HETATM   37  C29 UNK Z   1      24.200  29.015  30.245  1.00  0.00           C
HETATM   38  H1  UNK Z   1      21.885  26.761  44.308  1.00  0.00           H
HETATM   39  H2  UNK Z   1      21.945  26.364  41.883  1.00  0.00           H
HETATM   40  H3  UNK Z   1      26.244  25.937  42.090  1.00  0.00           H
HETATM   41  H4  UNK Z   1      29.563  28.445  42.468  1.00  0.00           H
HETATM   42  H5  UNK Z   1      31.364  27.539  43.957  1.00  0.00           H
HETATM   43  H6  UNK Z   1      32.530  27.119  45.821  1.00  0.00           H
HETATM   44  H7  UNK Z   1      34.051  25.878  47.349  1.00  0.00           H
HETATM   45  H8  UNK Z   1      33.205  23.805  48.481  1.00  0.00           H
HETATM   46  H9  UNK Z   1      29.520  24.136  46.688  1.00  0.00           H
HETATM   47  H10 UNK Z   1      26.019  23.814  37.672  1.00  0.00           H
HETATM   48  H11 UNK Z   1      25.317  23.508  35.323  1.00  0.00           H
HETATM   49  H12 UNK Z   1      22.188  26.432  35.873  1.00  0.00           H
HETATM   50  H13 UNK Z   1      22.883  26.754  38.211  1.00  0.00           H
HETATM   51  H14 UNK Z   1      24.624  26.168  46.633  1.00  0.00           H
HETATM   52  H15 UNK Z   1      24.170  27.850  46.328  1.00  0.00           H
HETATM   53  H16 UNK Z   1      22.934  26.589  46.459  1.00  0.00           H
HETATM   54  H17 UNK Z   1      26.311  25.841  45.590  1.00  0.00           H
HETATM   55  H18 UNK Z   1      23.072  25.860  40.022  1.00  0.00           H
HETATM   56  H19 UNK Z   1      22.280  24.429  33.965  1.00  0.00           H
HETATM   57  H20 UNK Z   1      23.923  24.049  33.497  1.00  0.00           H
HETATM   58  H21 UNK Z   1      21.814  25.448  31.968  1.00  0.00           H
HETATM   59  H22 UNK Z   1      23.432  25.077  31.410  1.00  0.00           H
HETATM   60  H23 UNK Z   1      22.135  27.874  31.592  1.00  0.00           H
HETATM   61  H24 UNK Z   1      22.415  26.989  30.106  1.00  0.00           H
HETATM   62  H25 UNK Z   1      24.745  27.096  30.485  1.00  0.00           H
HETATM   63  H26 UNK Z   1      24.242  28.633  32.938  1.00  0.00           H
HETATM   64  H27 UNK Z   1      25.807  28.223  32.261  1.00  0.00           H
HETATM   65  H28 UNK Z   1      25.353  25.789  32.569  1.00  0.00           H
HETATM   66  H29 UNK Z   1      25.201  26.658  34.084  1.00  0.00           H
HETATM   67  H30 UNK Z   1      25.221  29.389  30.153  1.00  0.00           H
HETATM   68  H31 UNK Z   1      23.814  28.860  29.237  1.00  0.00           H
HETATM   69  H32 UNK Z   1      23.597  29.784  30.730  1.00  0.00           H
TER
ENDMDL
```

Position restrained energy minimization file:

```
; VARIOUS PREPROCESSING OPTIONS
; Preprocessor information: use cpp syntax.
; e.g.: -I/home/joe/doe -I/home/mary/hoe
include                  =
; e.g.: -DI_Want_Cookies -DMe_Too
define                   = -DPOSRES -DPOSRES_LIGAND
```

```
; RUN CONTROL PARAMETERS
integrator               = steep
; Start time and timestep in ps
tinit                    = 0
dt                       = 0.001
nsteps                   = 20000
; For exact run continuation or redoing part of a run
; Part index is updated automatically on checkpointing (keeps files
separate)
simulation_part          = 1
init_step                = 0
; mode for center of mass motion removal
comm-mode                = Linear
; number of steps for center of mass motion removal
nstcomm                  = 1
; group(s) for center of mass motion removal
comm-grps                = Protein_LIG Water_and_ions

; LANGEVIN DYNAMICS OPTIONS
; Friction coefficient (amu/ps) and random seed
bd-fric                  = 0
ld-seed                  = 1993

; ENERGY MINIMIZATION OPTIONS
; Force tolerance and initial step-size
emtol                    = 10
emstep                   = 0.01
; Max number of iterations in relax_shells
niter                    = 20
; Step size (ps^2) for minimization of flexible constraints
fcstep                   = 0
; Frequency of steepest descents steps when doing CG
nstcgsteep               = 1000
nbfgscorr                = 10

; TEST PARTICLE INSERTION OPTIONS
rtpi                     = 0.05

; OUTPUT CONTROL OPTIONS
; Output frequency for coords (x), velocities (v) and forces (f)
nstxout                  = 1000
nstvout                  = 1000
nstfout                  = 1000
; Output frequency for energies to log file and energy file
nstlog                   = 100
nstenergy                = 100
; Output frequency and precision for xtc file
nstxtcout                = 100
xtc-precision            = 1000
; This selects the subset of atoms for the xtc file. You can
; select multiple groups. By default all atoms will be written.
xtc-grps                 =
; Selection of energy groups
energygrps               = Protein_LIG Water_and_ions

; NEIGHBORSEARCHING PARAMETERS
; nblist update frequency
nstlist                  = 5
; ns algorithm (simple or grid)
ns-type                  = Grid
```

```
; Periodic boundary conditions: xyz, no, xy
pbc                      = xyz
periodic_molecules       = no
; nblist cut-off
rlist                    = 1.0

; OPTIONS FOR ELECTROSTATICS AND VDW
; Method for doing electrostatics
coulombtype              = PME
rcoulomb-switch          = 0
rcoulomb                 = 1.0
; Relative dielectric constant for the medium and the reaction field
epsilon_r                = 1
epsilon_rf               = 66
; Method for doing Van der Waals
vdw-type                 = Cut-off
; cut-off lengths
rvdw-switch              = 0
rvdw                     = 1.0
; Apply long range dispersion corrections for Energy and Pressure
DispCorr                 = EnerPres
; Extension of the potential lookup tables beyond the cut-off
table-extension          = 1
; Seperate tables between energy group pairs
energygrp_table          =

; OPTIONS FOR WEAK COUPLING ALGORITHMS
; Temperature coupling
tcoupl                   = Berendsen
; Groups to couple separately
tc-grps                  = Protein_LIG Water_and_ions
; Time constant (ps) and reference temperature (K)
tau-t                    = 0.1 0.1
ref-t                    = 310 310
; Pressure coupling
Pcoupl                   = Berendsen
Pcoupltype               = Isotropic
; Time constant (ps), compressibility (1/bar) and reference P (bar)
tau-p                    = 1
compressibility          = 4.5e-5
ref-p                    = 1.0
; Scaling of reference coordinates, No, All or COM
refcoord_scaling         = No
; Random seed for Andersen thermostat
andersen_seed            = 815131

; SIMULATED ANNEALING
; Type of annealing for each temperature group (no/single/periodic)
annealing                =
; Number of time points to use for specifying annealing in each group
annealing_npoints        =
; List of times at the annealing points for each group
annealing_time           =
; Temp. at each annealing point, for each group.
annealing_temp           =

; GENERATE VELOCITIES FOR STARTUP RUN
gen-vel                  = no
gen-temp                 = 310
gen-seed                 = 173529
```

```
; OPTIONS FOR BONDS
constraints              = none
; Type of constraint algorithm
constraint-algorithm     = Lincs
; Do not constrain the start configuration
continuation             = no
; Use successive overrelaxation to reduce the number of shake iterations
Shake-SOR                = no
; Relative tolerance of shake
shake-tol                = 0.0001
; Highest order in the expansion of the constraint coupling matrix
lincs-order              = 4
; Number of iterations in the final step of LINCS. 1 is fine for
; normal simulations, but use 2 to conserve energy in NVE runs.
; For energy minimization with constraints it should be 4 to 8.
lincs-iter               = 1
; Lincs will write a warning to the stderr if in one step a bond
; rotates over more degrees than
lincs-warnangle          = 30
; Convert harmonic bonds to morse potentials
morse                    = no
```

Energy minimization file (free of restraints):

```
; VARIOUS PREPROCESSING OPTIONS
; Preprocessor information: use cpp syntax.
; e.g.: -I/home/joe/doe -I/home/mary/hoe
include                  =
; e.g.: -DI_Want_Cookies -DMe_Too
define                   =

; RUN CONTROL PARAMETERS
integrator               = steep
; Start time and timestep in ps
tinit                    = 0
dt                       = 0.001
nsteps                   = 20000
; For exact run continuation or redoing part of a run
; Part index is updated automatically on checkpointing (keeps files
separate)
simulation_part          = 1
init_step                = 0
; mode for center of mass motion removal
comm-mode                = Linear
; number of steps for center of mass motion removal
nstcomm                  = 1
; group(s) for center of mass motion removal
comm-grps                = Protein_LIG Water_and_ions

; LANGEVIN DYNAMICS OPTIONS
; Friction coefficient (amu/ps) and random seed
bd-fric                  = 0
ld-seed                  = 1993

; ENERGY MINIMIZATION OPTIONS
; Force tolerance and initial step-size
emtol                    = 10
```

```
emstep                   = 0.01
; Max number of iterations in relax_shells
niter                    = 20
; Step size (ps^2) for minimization of flexible constraints
fcstep                   = 0
; Frequency of steepest descents steps when doing CG
nstcgsteep               = 1000
nbfgscorr                = 10

; TEST PARTICLE INSERTION OPTIONS
rtpi                     = 0.05

; OUTPUT CONTROL OPTIONS
; Output frequency for coords (x), velocities (v) and forces (f)
nstxout                  = 1000
nstvout                  = 1000
nstfout                  = 1000
; Output frequency for energies to log file and energy file
nstlog                   = 100
nstenergy                = 100
; Output frequency and precision for xtc file
nstxtcout                = 100
xtc-precision            = 1000
; This selects the subset of atoms for the xtc file. You can
; select multiple groups. By default all atoms will be written.
xtc-grps                 =
; Selection of energy groups
energygrps               = Protein_LIG Water_and_ions

; NEIGHBORSEARCHING PARAMETERS
; nblist update frequency
nstlist                  = 5
; ns algorithm (simple or grid)
ns-type                  = Grid
; Periodic boundary conditions: xyz, no, xy
pbc                      = xyz
periodic_molecules       = no
; nblist cut-off
rlist                    = 1.0

; OPTIONS FOR ELECTROSTATICS AND VDW
; Method for doing electrostatics
coulombtype              = PME
rcoulomb-switch          = 0
rcoulomb                 = 1.0
; Relative dielectric constant for the medium and the reaction field
epsilon_r                = 1
epsilon_rf               = 66
; Method for doing Van der Waals
vdw-type                 = Cut-off
; cut-off lengths
rvdw-switch              = 0
rvdw                     = 1.4
; Apply long range dispersion corrections for Energy and Pressure
DispCorr                 = EnerPres
; Extension of the potential lookup tables beyond the cut-off
table-extension          = 1
; Seperate tables between energy group pairs
energygrp_table          =

; OPTIONS FOR WEAK COUPLING ALGORITHMS
```

```
; Temperature coupling
tcoupl                  = Berendsen
; Groups to couple separately
tc-grps                 = Protein_LIG Water_and_ions
; Time constant (ps) and reference temperature (K)
tau-t                   = 0.1 0.1
ref-t                   = 310 310
; Pressure coupling
Pcoupl                  = Berendsen
Pcoupltype              = Isotropic
; Time constant (ps), compressibility (1/bar) and reference P (bar)
tau-p                   = 1
compressibility         = 4.5e-5
ref-p                   = 1.0
; Scaling of reference coordinates, No, All or COM
refcoord_scaling        = No
; Random seed for Andersen thermostat
andersen_seed           = 815131

; SIMULATED ANNEALING
; Type of annealing for each temperature group (no/single/periodic)
annealing               =
; Number of time points to use for specifying annealing in each group
annealing_npoints       =
; List of times at the annealing points for each group
annealing_time          =
; Temp. at each annealing point, for each group.
annealing_temp          =

; GENERATE VELOCITIES FOR STARTUP RUN
gen-vel                 = no
gen-temp                = 310
gen-seed                = 173529

; OPTIONS FOR BONDS
constraints             = none
; Type of constraint algorithm
constraint-algorithm    = Lincs
; Do not constrain the start configuration
continuation            = no
; Use successive overrelaxation to reduce the number of shake iterations
Shake-SOR               = no
; Relative tolerance of shake
shake-tol               = 0.0001
; Highest order in the expansion of the constraint coupling matrix
lincs-order             = 4
; Number of iterations in the final step of LINCS. 1 is fine for
; normal simulations, but use 2 to conserve energy in NVE runs.
; For energy minimization with constraints it should be 4 to 8.
lincs-iter              = 1
; Lincs will write a warning to the stderr if in one step a bond
; rotates over more degrees than
lincs-warnangle         = 30
; Convert harmonic bonds to morse potentials
morse                   = no
```

Position-restrained MD:

```
; VARIOUS PREPROCESSING OPTIONS
; Preprocessor information: use cpp syntax.
; e.g.: -I/home/joe/doe -I/home/mary/hoe
include                  =
; e.g.: -DI_Want_Cookies -DMe_Too
define                   = -DPOSRES

; RUN CONTROL PARAMETERS
integrator               = md
; Start time and timestep in ps
tinit                    = 0
dt                       = 0.002
nsteps                   = 250000
; For exact run continuation or redoing part of a run
; Part index is updated automatically on checkpointing (keeps files
separate)
simulation_part          = 1
init_step                = 0
; mode for center of mass motion removal
comm-mode                = Linear
; number of steps for center of mass motion removal
nstcomm                  = 1
; group(s) for center of mass motion removal
comm-grps                = Protein_LIG Water_and_ions

; LANGEVIN DYNAMICS OPTIONS
; Friction coefficient (amu/ps) and random seed
bd-fric                  = 0
ld-seed                  = 1993

; ENERGY MINIMIZATION OPTIONS
; Force tolerance and initial step-size
emtol                    = 10
emstep                   = 0.01
; Max number of iterations in relax_shells
niter                    = 20
; Step size (ps^2) for minimization of flexible constraints
fcstep                   = 0
; Frequency of steepest descents steps when doing CG
nstcgsteep               = 1000
nbfgscorr                = 10

; TEST PARTICLE INSERTION OPTIONS
rtpi                     = 0.05

; OUTPUT CONTROL OPTIONS
; Output frequency for coords (x), velocities (v) and forces (f)
nstxout                  = 5000
nstvout                  = 5000
nstfout                  = 5000
; Output frequency for energies to log file and energy file
nstlog                   = 100
nstenergy                = 100
; Output frequency and precision for xtc file
nstxtcout                = 500
xtc-precision            = 1000
; This selects the subset of atoms for the xtc file. You can
; select multiple groups. By default all atoms will be written.
xtc-grps                 =
; Selection of energy groups
```

```
energygrps               = Protein_LIG Water_and_ions

; NEIGHBORSEARCHING PARAMETERS
; nblist update frequency
nstlist                  = 5
; ns algorithm (simple or grid)
ns-type                  = Grid
; Periodic boundary conditions: xyz, no, xy
pbc                      = xyz
periodic_molecules       = no
; nblist cut-off
rlist                    = 1

; OPTIONS FOR ELECTROSTATICS AND VDW
; Method for doing electrostatics
coulombtype              = PME
rcoulomb-switch          = 0
rcoulomb                 = 1.0
; Relative dielectric constant for the medium and the reaction field
epsilon_r                = 1
epsilon_rf               = 66
; Method for doing Van der Waals
vdw-type                 = Cut-off
; cut-off lengths
rvdw-switch              = 0
rvdw                     = 1.4
; Apply long range dispersion corrections for Energy and Pressure
DispCorr                 = EnerPres
; Extension of the potential lookup tables beyond the cut-off
table-extension          = 1
; Seperate tables between energy group pairs
energygrp_table          =
; Spacing for the PME/PPPM FFT grid
fourierspacing           = 0.12
; FFT grid size, when a value is 0 fourierspacing will be used
fourier_nx               = 0
fourier_ny               = 0
fourier_nz               = 0
; EWALD/PME/PPPM parameters
pme_order                = 4
ewald_rtol               = 1e-05
ewald_geometry           = 3d
epsilon_surface          = 0
optimize_fft             = no

; IMPLICIT SOLVENT ALGORITHM
implicit_solvent         = No

; GENERALIZED BORN ELECTROSTATICS
; Algorithm for calculating Born radii
gb_algorithm             = Still
; Frequency of calculating the Born radii inside rlist
nstgbradii               = 1
; Cutoff for Born radii calculation; the contribution from atoms
; between rlist and rgbradii is updated every nstlist steps
rgbradii                 = 2
; Dielectric coefficient of the implicit solvent
gb_epsilon_solvent       = 80
; Salt concentration in M for Generalized Born models
gb_saltconc              = 0
; Scaling factors used in the OBC GB model. Default values are OBC(II)
```

```
gb_obc_alpha             = 1
gb_obc_beta              = 0.8
gb_obc_gamma             = 4.85
; Surface tension (kJ/mol/nm^2) for the SA (nonpolar surface) part of GBSA
; The default value (2.092) corresponds to 0.005 kcal/mol/Angstrom^2.
sa_surface_tension       = 2.092

; OPTIONS FOR WEAK COUPLING ALGORITHMS
; Temperature coupling
tcoupl                   = Berendsen
; Groups to couple separately
tc-grps                  = Protein_LIG Water_and_ions
; Time constant (ps) and reference temperature (K)
tau-t                    = 0.1 0.1
ref-t                    = 310 310
; Pressure coupling
Pcoupl                   = ;Berendsen
Pcoupltype               = ;Isotropic
; Time constant (ps), compressibility (1/bar) and reference P (bar)
tau-p                    = 1
compressibility          = 4.5e-5
ref-p                    = 1.0
; Scaling of reference coordinates, No, All or COM
refcoord_scaling         = No
; Random seed for Andersen thermostat
andersen_seed            = 815131

; OPTIONS FOR QMMM calculations
QMMM                     = no
; Groups treated Quantum Mechanically
QMMM-grps                =
; QM method
QMmethod                 =
; QMMM scheme
QMMMscheme               = normal
; QM basisset
QMbasis                  =
; QM charge
QMcharge                 =
; QM multiplicity
QMmult                   =
; Surface Hopping
SH                       =
; CAS space options
CASorbitals              =
CASelectrons             =
SAon                     =
SAoff                    =
SAsteps                  =
; Scale factor for MM charges
MMChargeScaleFactor      = 1
; Optimization of QM subsystem
bOPT                     =
bTS                      =

; SIMULATED ANNEALING
; Type of annealing for each temperature group (no/single/periodic)
annealing                =
; Number of time points to use for specifying annealing in each group
annealing_npoints        =
; List of times at the annealing points for each group
```

```
annealing_time           =
; Temp. at each annealing point, for each group.
annealing_temp           =

; GENERATE VELOCITIES FOR STARTUP RUN
gen-vel                  = no
gen-temp                 = 310
gen-seed                 = 173529

; OPTIONS FOR BONDS
constraints              = all-bonds
; Type of constraint algorithm
constraint-algorithm     = Lincs
; Do not constrain the start configuration
continuation             = yes
; Use successive overrelaxation to reduce the number of shake iterations
Shake-SOR                = no
; Relative tolerance of shake
shake-tol                = 0.0001
; Highest order in the expansion of the constraint coupling matrix
lincs-order              = 4
; Number of iterations in the final step of LINCS. 1 is fine for
; normal simulations, but use 2 to conserve energy in NVE runs.
; For energy minimization with constraints it should be 4 to 8.
lincs-iter               = 1
; Lincs will write a warning to the stderr if in one step a bond
; rotates over more degrees than
lincs-warnangle          = 30
; Convert harmonic bonds to morse potentials
morse                    = no

; ENERGY GROUP EXCLUSIONS
; Pairs of energy groups for which all non-bonded interactions are excluded
energygrp_excl           =

; WALLS
; Number of walls, type, atom types, densities and box-z scale factor for
Ewald
nwall                    = 0
wall_type                = 9-3
wall_r_linpot            = -1
wall_atomtype            =
wall_density             =
wall_ewald_zfac          = 3

; COM PULLING
; Pull type: no, umbrella, constraint or constant_force
pull                     = no

; NMR refinement stuff
; Distance restraints type: No, Simple or Ensemble
disre                    = No
; Force weighting of pairs in one distance restraint: Conservative or Equal
disre-weighting          = Conservative
; Use sqrt of the time averaged times the instantaneous violation
disre-mixed              = no
disre-fc                 = 1000
disre-tau                = 0
; Output frequency for pair distances to energy file
nstdisreout              = 100
; Orientation restraints: No or Yes
```

```
orire                    = no
; Orientation restraints force constant and tau for time averaging
orire-fc                 = 0
orire-tau                = 0
orire-fitgrp             =
; Output frequency for trace(SD) and S to energy file
nstorireout              = 100
; Dihedral angle restraints: No or Yes
dihre                    = no
dihre-fc                 = 1000

; Free energy control stuff
free-energy              = no
init-lambda              = 0
delta-lambda             = 0
sc-alpha                 = 0
sc-power                 = 0
sc-sigma                 = 0.3
couple-moltype           =
couple-lambda0           = vdw-q
couple-lambda1           = vdw-q
couple-intramol          = no

; Non-equilibrium MD stuff
acc-grps                 =
accelerate               =
freezegrps               =
freezedim                =
cos-acceleration         = 0
deform                   =

; Electric fields
; Format is number of terms (int) and for all terms an amplitude (real)
; and a phase angle (real)
E-x                      =
E-xt                     =
E-y                      =
E-yt                     =
E-z                      =
E-zt                     =

; User defined thingies
user1-grps               =
user2-grps               =
userint1                 = 0
userint2                 = 0
userint3                 = 0
userint4                 = 0
userreal1                = 0
userreal2                = 0
userreal3                = 0
userreal4                = 0
```

Production MD runs:

```
; VARIOUS PREPROCESSING OPTIONS
; Preprocessor information: use cpp syntax.
; e.g.: -I/home/joe/doe -I/home/mary/hoe
```

```
include                  =
; e.g.: -DI_Want_Cookies -DMe_Too
define                   =

; RUN CONTROL PARAMETERS
integrator               = md
; Start time and timestep in ps
tinit                    = 0
dt                       = 0.002
nsteps                   = 10000000
; For exact run continuation or redoing part of a run
; Part index is updated automatically on checkpointing (keeps files
separate)
simulation_part          = 1
init_step                = 0
; mode for center of mass motion removal
comm-mode                = Linear
; number of steps for center of mass motion removal
nstcomm                  = 1
; group(s) for center of mass motion removal
comm-grps                = Protein_LIG Water_and_ions

; LANGEVIN DYNAMICS OPTIONS
; Friction coefficient (amu/ps) and random seed
bd-fric                  = 0
ld-seed                  = 1993

; ENERGY MINIMIZATION OPTIONS
; Force tolerance and initial step-size
emtol                    = 10
emstep                   = 0.01
; Max number of iterations in relax_shells
niter                    = 20
; Step size (ps^2) for minimization of flexible constraints
fcstep                   = 0
; Frequency of steepest descents steps when doing CG
nstcgsteep               = 1000
nbfgscorr                = 10

; TEST PARTICLE INSERTION OPTIONS
rtpi                     = 0.05

; OUTPUT CONTROL OPTIONS
; Output frequency for coords (x), velocities (v) and forces (f)
nstxout                  = 50000
nstvout                  = 50000
nstfout                  = 50000
; Output frequency for energies to log file and energy file
nstlog                   = 100
nstenergy                = 100
; Output frequency and precision for xtc file
nstxtcout                = 500
xtc-precision            = 1000
; This selects the subset of atoms for the xtc file. You can
; select multiple groups. By default all atoms will be written.
xtc-grps                 =
; Selection of energy groups
energygrps               = Protein_LIG Water_and_ions

; NEIGHBORSEARCHING PARAMETERS
; nblist update frequency
```

```
nstlist                  = 5
; ns algorithm (simple or grid)
ns-type                  = Grid
; Periodic boundary conditions: xyz, no, xy
pbc                      = xyz
periodic_molecules       = no
; nblist cut-off
rlist                    = 0.90

; OPTIONS FOR ELECTROSTATICS AND VDW
; Method for doing electrostatics
coulombtype              = PME
rcoulomb-switch          = 0
rcoulomb                 = 0.90
; Relative dielectric constant for the medium and the reaction field
epsilon_r                = 1
epsilon_rf               = 66
; Method for doing Van der Waals
vdw-type                 = Cut-off
; cut-off lengths
rvdw-switch              = 0
rvdw                     = 1.4
; Apply long range dispersion corrections for Energy and Pressure
DispCorr                 = EnerPres
; Extension of the potential lookup tables beyond the cut-off
table-extension          = 1
; Seperate tables between energy group pairs
energygrp_table          =
; Spacing for the PME/PPPM FFT grid
fourierspacing           = 0.12
; FFT grid size, when a value is 0 fourierspacing will be used
fourier_nx               = 0
fourier_ny               = 0
fourier_nz               = 0
; EWALD/PME/PPPM parameters
pme_order                = 4
ewald_rtol               = 1e-05
ewald_geometry           = 3d
epsilon_surface          = 0
optimize_fft             = no

; IMPLICIT SOLVENT ALGORITHM
implicit_solvent         = No

; GENERALIZED BORN ELECTROSTATICS
; Algorithm for calculating Born radii
gb_algorithm             = Still
; Frequency of calculating the Born radii inside rlist
nstgbradii               = 1
; Cutoff for Born radii calculation; the contribution from atoms
; between rlist and rgbradii is updated every nstlist steps
rgbradii                 = 2
; Dielectric coefficient of the implicit solvent
gb_epsilon_solvent       = 80
; Salt concentration in M for Generalized Born models
gb_saltconc              = 0
; Scaling factors used in the OBC GB model. Default values are OBC(II)
gb_obc_alpha             = 1
gb_obc_beta              = 0.8
gb_obc_gamma             = 4.85
; Surface tension (kJ/mol/nm^2) for the SA (nonpolar surface) part of GBSA
```

```
; The default value (2.092) corresponds to 0.005 kcal/mol/Angstrom^2.
sa_surface_tension       = 2.092

; OPTIONS FOR WEAK COUPLING ALGORITHMS
; Temperature coupling
tcoupl                   = v-rescale
; Groups to couple separately
tc-grps                  = Protein_LIG Water_and_ions
; Time constant (ps) and reference temperature (K)
tau-t                    = 0.1 0.1
ref-t                    = 310 310
; Pressure coupling
Pcoupl                   = Berendsen
Pcoupltype               = Isotropic
; Time constant (ps), compressibility (1/bar) and reference P (bar)
tau-p                    = 1
compressibility          = 4.5e-5
ref-p                    = 1.0
; Scaling of reference coordinates, No, All or COM
refcoord_scaling         = No
; Random seed for Andersen thermostat
andersen_seed            = 815131

; OPTIONS FOR QMMM calculations
QMMM                     = no
; Groups treated Quantum Mechanically
QMMM-grps                =
; QM method
QMmethod                 =
; QMMM scheme
QMMMscheme               = normal
; QM basisset
QMbasis                  =
; QM charge
QMcharge                 =
; QM multiplicity
QMmult                   =
; Surface Hopping
SH                       =
; CAS space options
CASorbitals              =
CASelectrons             =
SAon                     =
SAoff                    =
SAsteps                  =
; Scale factor for MM charges
MMChargeScaleFactor      = 1
; Optimization of QM subsystem
bOPT                     =
bTS                      =

; SIMULATED ANNEALING
; Type of annealing for each temperature group (no/single/periodic)
annealing                =
; Number of time points to use for specifying annealing in each group
annealing_npoints        =
; List of times at the annealing points for each group
annealing_time           =
; Temp. at each annealing point, for each group.
annealing_temp           =
```

```
; GENERATE VELOCITIES FOR STARTUP RUN
gen-vel                  = yes
gen-temp                 = 10
gen-seed                 = -1

; OPTIONS FOR BONDS
constraints              = all-bonds
; Type of constraint algorithm
constraint-algorithm     = Lincs
; Do not constrain the start configuration
continuation             = no
; Use successive overrelaxation to reduce the number of shake iterations
Shake-SOR                = no
; Relative tolerance of shake
shake-tol                = 0.0001
; Highest order in the expansion of the constraint coupling matrix
lincs-order              = 4
; Number of iterations in the final step of LINCS. 1 is fine for
; normal simulations, but use 2 to conserve energy in NVE runs.
; For energy minimization with constraints it should be 4 to 8.
lincs-iter               = 1
; Lincs will write a warning to the stderr if in one step a bond
; rotates over more degrees than
lincs-warnangle          = 30
; Convert harmonic bonds to morse potentials
morse                    = no

; ENERGY GROUP EXCLUSIONS
; Pairs of energy groups for which all non-bonded interactions are excluded
energygrp_excl           =

; WALLS
; Number of walls, type, atom types, densities and box-z scale factor for
Ewald
nwall                    = 0
wall_type                = 9-3
wall_r_linpot            = -1
wall_atomtype            =
wall_density             =
wall_ewald_zfac          = 3

; COM PULLING
; Pull type: no, umbrella, constraint or constant_force
pull                     = no

; NMR refinement stuff
; Distance restraints type: No, Simple or Ensemble
disre                    = No
; Force weighting of pairs in one distance restraint: Conservative or Equal
disre-weighting          = Conservative
; Use sqrt of the time averaged times the instantaneous violation
disre-mixed              = no
disre-fc                 = 1000
disre-tau                = 0
; Output frequency for pair distances to energy file
nstdisreout              = 100
; Orientation restraints: No or Yes
orire                    = no
; Orientation restraints force constant and tau for time averaging
orire-fc                 = 0
orire-tau                = 0
```

```
orire-fitgrp             =
; Output frequency for trace(SD) and S to energy file
nstorireout              = 100
; Dihedral angle restraints: No or Yes
dihre                    = no
dihre-fc                 = 1000

; Free energy control stuff
free-energy              = no
init-lambda              = 0
delta-lambda             = 0
sc-alpha                 = 0
sc-power                 = 0
sc-sigma                 = 0.3
couple-moltype           =
couple-lambda0           = vdw-q
couple-lambda1           = vdw-q
couple-intramol          = no

; Non-equilibrium MD stuff
acc-grps                 =
accelerate               =
freezegrps               =
freezedim                =
cos-acceleration         = 0
deform                   =

; Electric fields
; Format is number of terms (int) and for all terms an amplitude (real)
; and a phase angle (real)
E-x                      =
E-xt                     =
E-y                      =
E-yt                     =
E-z                      =
E-zt                     =

; User defined thingies
user1-grps               =
user2-grps               =
userint1                 = 0
userint2                 = 0
userint3                 = 0
userint4                 = 0
userreal1                = 0
userreal2                = 0
userreal3                = 0
userreal4                = 0
```

2.3. Energy analysis

MM-PBSA parameter file:

```
;Polar calculation: "yes" or "no"
polar       = yes
;=============
;PSIZE options
;=============
;Factor by which to expand molecular dimensions to get coarsegrid
dimensions.
cfac        = 1.5
;The desired fine mesh spacing (in A)
gridspace   = 0.5
:Amount (in A) to add to molecular dimensions to get fine grid dimensions.
fadd        = 5
;Maximum memory (in MB) available per-processor for a calculation.
gmemceil    = 4000
;============================================
;APBS kwywords for polar solvation calculation
;============================================
;Charge of positive ions
pcharge     = 1
;Radius of positive charged ions
prad        = 0.95
;Concentration of positive charged ions
pconc           = 0.150
;Charge of negative ions
ncharge     = -1
;Radius of negative charged ions
nrad        = 1.81
;Concentration of negative charged ions
nconc             = 0.150
;Solute dielectric constant
pdie        = 2
;Solvent dielectric constant
sdie        = 80
;Reference or vacuum dielectric constant
vdie        = 1
;Solvent probe radius
srad        = 1.4
;Method used to map biomolecular charges on grid. chgm = spl0 or spl2 or
spl4
chgm            = spl4
;Model used to construct dielectric and ionic boundary. srfm = smol or spl2
or spl4
srfm            = smol
;Value for cubic spline window. Only used in case of srfm = spl2 or spl4.
swin        = 0.30
;Numebr of grid point per A^2. Not used when (srad = 0.0) or (srfm = spl2
or spl4)
sdens             = 10
;Temperature in K
temp        = 300
;Type of boundary condition to solve PB equation. bcfl = zero or sdh or mdh
or focus or map
bcfl        = mdh
;Non-linear (npbe) or linear (lpbe) PB equation to solve
PBsolver    = lpbe
```

```
;============================================================
;APBS kwywords for Apolar/Non-polar solvation calculation
;============================================================
;Non-polar solvation calculation: "yes" or "no"
apolar            = yes
;Repulsive contribution to Non-polar
;===SASA model ====
;Gamma (Surface Tension) kJ/(mol A^2)
gamma             = 0.0226778
;Probe radius for SASA (A)
sasrad            = 1.4
;Offset (c) kJ/mol
sasaconst         = 3.84928
;===SAV model===
;Pressure kJ/(mol A^3)
press             = 0
;Probe radius for SAV (A)
savrad            = 0
;Offset (c) kJ/mol
savconst          = 0
;Attractive contribution to Non-polar
;===WCA model ====
;using WCA method: "yes" or "no"
WCA               = no
;Probe radius for WCA
wcarad            = 1.20
;bulk solvent density in A^3
bconc       = 0.033428
;displacment in A for surface area derivative calculation
dpos        = 0.05
;Quadrature grid points per A for molecular surface or solvent accessible
surface
APsdens            = 20
;Quadrature grid spacing in A for volume integral calculations
grid              = 0.45 0.45 0.45
;Parameter to construct solvent related surface or volume
APsrfm            = sacc
;Cubic spline window in A for spline based surface definitions
APswin            = 0.3
;Temperature in K
APtemp            = 300
```

Bibliography

ADAMS, J. A. **Kinetic and catalytic mechanisms of protein kinases.** *Chemical reviews*, v. 101, n. 8, p. 2271–90, 2001.

ALEKSANDROV, A.; SIMONSON, T. **A molecular mechanics model for imatinib and imatinib:kinase binding.** *Journal of computational chemistry*, v. 31, n. 7, p. 1550–60, 2010.

ALLAIN, A. et al. **Allosteric pathway identification through network analysis: from molecular dynamics simulations to interactive 2D and 3D graphs.** *Faraday discussions*, v. 169, p. 303–21, 2014.

ALLEN, M. P.; TILDESLEY, D. J. **Computer Simulation of Liquids**. New York, NY, USA: Clarendon Press, 1989.

ALTSCHUL, S. **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Research*, v. 25, n. 17, p. 3389–3402, 1997.

ALTSCHUL, S. F. et al. **Basic local alignment search tool.** *Journal of molecular biology*, v. 215, n. 3, p. 403–10, 1990.

AMADEI, A.; LINSSEN, A. B. M.; BERENDSEN, H. J. C. **Essential Dynamics of Proteins**. *Proteins-Structure Function and Genetics*, v. 17, n. 4, p. 412–425, 1993.

ANFINSEN, C. B. **Principles that govern the folding of protein chains.** *Science (New York, N.Y.)*, v. 181, n. 4096, p. 223–30, 1973.

ASAI, K.; HAYAMIZU, S.; HANDA, K. **Prediction of protein secondary structure by the hidden Markov model.** *Computer applications in the biosciences : CABIOS*, v. 9, n. 2, p. 141–6, 1993.

ASHMAN, L. K. et al. **Effects of mutant c-kit in early myeloid cells.** *Leukemia & lymphoma*, v. 37, n. 1-2, p. 233–43, 2000.

AZAM, M. et al. **Activation of tyrosine kinases by mutation of the gatekeeper threonine.** *Nature structural & molecular biology*, v. 15, n. 10, p. 1109–18, 2008.

BAHAR, I. et al. **Normal Mode Analysis of Biomolecular Structures: Functional Mechanisms of Membrane Proteins**. *Chemical Reviews*, v. 110, n. 3, p. 1463–1497, 2010.

BAKER, D.; SALI, A. **Protein structure prediction and structural genomics.** *Science (New York, N.Y.)*, v. 294, n. 5540, p. 93–6, 2001.

BAKER, N. A. et al. **Electrostatics of nanosystems: Application to microtubules and the ribosome**. *Proceedings of the National Academy of Sciences of the United States of America*, v. 98, n. 18, p. 10037–10041, 2001.

BASHFORD, D.; CASE, D. A. **Generalized born models of macromolecular solvation effects.** *Annual review of physical chemistry*, v. 51, p. 129–52, 2000.

BATISTA, P. R. et al. **Consensus modes, a robust description of protein collective motions from multiple-minima normal mode analysis--application to the HIV-1 protease.** *Physical chemistry chemical physics : PCCP*, v. 12, n. 12, p. 2850–9, 2010.

BEGHINI, A. et al. **c-kit activating mutations and mast cell proliferation in human leukemia.** *Blood*, v. 92, n. 2, p. 701–2, 1998.

BENKERT, P.; BIASINI, M.; SCHWEDE, T. **Toward the estimation of the absolute quality of individual protein structure models.** *Bioinformatics (Oxford, England)*, v. 27, n. 3, p. 343–50, 2011.

BERENDSEN, H. **Collective protein dynamics in relation to function**. *Current Opinion in Structural Biology*, v. 10, n. 2, p. 165–169, 2000.

BERENDSEN, H. J. C. et al. **Molecular dynamics with coupling to an external bath**. *The Journal of Chemical Physics*, v. 81, n. 8, p. 3684, 1984.

BERMAN, H. M. et al. **The Protein Data Bank**. *Nucleic Acids Research*, v. 28, n. 1, p. 235–242, 2000.

BERMAN, H. M. **The Protein Data Bank**. *Nucleic Acids Research*, v. 28, n. 1, p. 235–242, 2000.

BISCHOF, R. J. et al. **Exacerbation of acute inflammatory arthritis by the colony-stimulating factors CSF-1 and granulocyte macrophage (GM)-CSF: evidence of macrophage infiltration and local proliferation**. *Clinical and Experimental Immunology*, v. 119, n. 2, p. 361–367, 2000.

BISHOP, A. O. T.; DE BEER, T. A. P.; JOUBERT, F. **Protein homology modelling and its use in South Africa**. *South African Journal of Science*, v. 104, n. 1-2, p. 2–6, 2008.

BIXON, M.; LIFSON, S. **Potential functions and conformations in cycloalkanes**. *Tetrahedron*, v. 23, n. 2, p. 769–784, 1967.

BLASZCZYK, M. et al. **Protein Structure Prediction Using CABS - A Consensus Approach**From Computational Biophysics to Systems Biology (CBSB11) Proceedings. **Anais**...2012

BLASZCZYK, M. et al. **CABS-fold: Server for the de novo and consensus-based prediction of protein structure.** *Nucleic acids research*, v. 41, n. Web Server issue, p. W406–11, 2013.

BLUME-JENSEN, P.; HUNTER, T. **Oncogenic kinase signalling.** *Nature*, v. 411, n. 6835, p. 355–65, 2001.

BLUNDELL, T. L. et al. **Knowledge-based prediction of protein structures and the design of novel molecules.** *Nature*, v. 326, n. 6111, p. 347–52, 1987.

BOUGHERARA, H. et al. **The Aberrant Localization of Oncogenic Kit Tyrosine Kinase Receptor Mutants Is Reversed on Specific Inhibitory Treatment**. *Molecular Cancer Research*, v. 7, n. 9, p. 1525–1533, 2009.

BOWIE, J. U.; LUTHY, R.; EISENBERG, D. **A Method to Identify Protein Sequences That Fold into a Known 3-Dimensional Structure**. *Science*, v. 253, n. 5016, p. 164–170, 1991.

BRADLEY, P.; MISURA, K. M. S.; BAKER, D. **Toward high-resolution de novo structure prediction for small proteins.** *Science (New York, N.Y.)*, v. 309, n. 5742, p. 1868–71, 2005.

BROOIJMANS, N.; KUNTZ, I. D. **Molecular recognition and docking algorithms**. *Annual Review of Biophysics and Biomolecular Structure*, v. 32, p. 335–373, 2003.

BROOKS, B.; KARPLUS, M. **Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor.** *Proceedings of the National Academy of Sciences*, v. 80, n. 21, p. 6571–6575, 1983.

BROOKS, B. R. et al. **CHARMM: A program for macromolecular energy, minimization, and dynamics calculations**. *Journal of Computational Chemistry*, v. 4, n. 2, p. 187–217, 1983.

BROOKS, B. R. et al. **CHARMM: the biomolecular simulation program.** *Journal of computational chemistry*, v. 30, n. 10, p. 1545–614, 2009.

BROWN, D.; CLARKE, J. H. R. **A comparison of constant energy, constant temperature and constant pressure ensembles in molecular dynamics simulations of atomic liquids**. *Molecular Physics*, v. 51, n. 5, p. 1243–1252, 2006.

BROWN, S. P.; MUCHMORE, S. W. **Large-scale application of high-throughput molecular mechanics with Poisson-Boltzmann surface area for routine physics-based scoring of protein-ligand complexes.** *Journal of medicinal chemistry*, v. 52, n. 10, p. 3159–65, 2009.

BRÜSCHWEILER, R. **Collective protein dynamics and nuclear spin relaxation**. *The Journal of Chemical Physics*, v. 102, n. 8, p. 3396, 1995.

BUSSI, G.; DONADIO, D.; PARRINELLO, M. **Canonical sampling through velocity rescaling.** *The Journal of chemical physics*, v. 126, n. 1, p. 014101, 2007.

CASE, D. A. et al. **The Amber biomolecular simulation programs.** *Journal of computational chemistry*, v. 26, n. 16, p. 1668–88, 2005.

CHANGEUX, J. P. **The feedback control mechanisms of biosynthetic L-threonine deaminase by L-isoleucine.** *Cold Spring Harbor symposia on quantitative biology*, v. 26, p. 313–8, 1961.

CHAUVOT DE BEAUCHÊNE, I. **Étude par modélisation moleculaire des mécanismes d'activation et de résistance du récepteur tyrosine kinase KIT sauvage et mutant**. [s.l.] École Normale Supérieure de Cachan, 2013.

CHAUVOT DE BEAUCHÊNE, I. et al. **Hotspot Mutations in KIT Receptor Differentially Modulate Its Allosterically Coupled Conformational Dynamics: Impact on Activation and Drug Sensitivity.** *PLoS computational biology*, v. 10, n. 7, p. e1003749, 2014.

CHEATHAM, T. E. et al. **Molecular-Dynamics Simulations on Solvated Biomolecular Systems - the Particle Mesh Ewald Method Leads to Stable Trajectories of DNA, Rna, and Proteins**. *Journal of the American Chemical Society*, v. 117, n. 14, p. 4193–4194, 1995.

CHENG, A.; MERZ, K. M. **Application of the Nosé−Hoover Chain Algorithm to the Study of Protein Dynamics**. *The Journal of Physical Chemistry*, v. 100, n. 5, p. 1927–1937, 1996.

CHENG, J. et al. **SCRATCH: a protein structure and structural feature prediction server.** *Nucleic acids research*, v. 33, n. Web Server issue, p. W72–6, 2005.

CHENNUBHOTLA, C.; BAHAR, I. **Markov propagation of allosteric effects in biomolecular systems: application to GroEL-GroES.** *Molecular systems biology*, v. 2, p. 36, 2006.

CHENNUBHOTLA, C.; BAHAR, I. **Signal propagation in proteins and relation to equilibrium fluctuations.** *PLoS computational biology*, v. 3, n. 9, p. 1716–26, 2007.

CHENNUBHOTLA, C.; YANG, Z.; BAHAR, I. **Coupling between global dynamics and signal transduction pathways: a mechanism of allostery for chaperonin GroEL.** *Molecular bioSystems*, v. 4, n. 4, p. 287–92, 2008.

CLINTON, S. K. et al. **Macrophage colony-stimulating factor gene expression in vascular cells and in experimental and human atherosclerosis.** *The American journal of pathology*, v. 140, n. 2, p. 301–16, 1992.

CLORE, G. M. et al. **Application of molecular dynamics with interproton distance restraints to three-dimensional protein structure determination. A model study of crambin.** *Journal of molecular biology*, v. 191, n. 3, p. 523–51, 1986.

COLE, C.; BARBER, J. D.; BARTON, G. J. **The Jpred 3 secondary structure prediction server.** *Nucleic acids research*, v. 36, n. Web Server issue, p. W197–201, 2008.

COLUMBO, M. et al. **The human recombinant c-kit receptor ligand, rhSCF, induces mediator release from human cutaneous mast cells and enhances IgE-dependent mediator release from both skin mast cells and peripheral blood basophils.** *Journal of immunology (Baltimore, Md. : 1950)*, v. 149, n. 2, p. 599–608, 1992.

CORLESS, C. L. et al. **PDGFRA mutations in gastrointestinal stromal tumors: frequency, spectrum and in vitro sensitivity to imatinib.** *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, v. 23, n. 23, p. 5357–64, 2005.

CUFF, J. A.; BARTON, G. J. **Application of multiple sequence alignment profiles to improve protein secondary structure prediction.** *Proteins*, v. 40, n. 3, p. 502–11, 2000.

DA SILVA FIGUEIREDO CELESTINO GOMES, P. et al. **Differential Effects of CSF-1R D802V and KIT D816V Homologous Mutations on Receptor Tertiary Structure and Allosteric Communication.** *PloS one*, v. 9, n. 5, p. e97519, 2014.

DA SILVA, M. L.; BISCH, P. M. **Modelagem molecular de proteínas da bactéria endofítica Gluconacetobacter diazotrophicus: Análise em larga escala e de proteínas potencialmente envolvidas na associação planta-bactéria**. Rio de Janeiro: UFRJ, 2011.

DARDEN, T.; YORK, D.; PEDERSEN, L. **Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems**. *The Journal of Chemical Physics*, v. 98, n. 12, p. 10089, 1993.

DEBIEC-RYCHTER, M. et al. **KIT mutations and dose selection for imatinib in patients with advanced gastrointestinal stromal tumours.** *European journal of cancer (Oxford, England : 1990)*, v. 42, n. 8, p. 1093–103, 2006.

DEL SOL, A. et al. **The origin of allosteric functional modulation: multiple pre-existing pathways.** *Structure (London, England : 1993)*, v. 17, n. 8, p. 1042–50, 2009.

DELANO, W. L. **Use of PYMOL as a communications tool for molecular science.** *Abstracts of Papers of the American Chemical Society*, v. 228, p. U313–U314, 2004.

DELANO, W. L.; LAM, J. W. **PyMOL: A communications tool for computational models**. *Abstracts of Papers of the American Chemical Society*, v. 230, p. U1371–U1372, 2005.

DEMETRI, G. D. et al. **Efficacy and Safety of Imatinib Mesylate in Advanced Gastrointestinal Stromal Tumors**. *New England Journal of Medicine*, v. 347, n. 7, p. 472–480, 2002.

DIBB, N. J.; DILWORTH, S. M.; MOL, C. D. **Switching on kinases: oncogenic activation of BRAF and the PDGFR family.** *Nature reviews. Cancer*, v. 4, n. 9, p. 718–27, 2004.

DINOLA, A.; ROCCATANO, D.; BERENDSEN, H. J. C. **Molecular-Dynamics Simulation of the Docking of Substrates to Proteins**. *Proteins-Structure Function and Genetics*, v. 19, n. 3, p. 174–182, 1994.

DIXIT, A.; VERKHIVKER, G. M. **Computational modeling of allosteric communication reveals organizing principles of mutation-induced signaling in ABL and EGFR kinases.** *PLoS computational biology*, v. 7, n. 10, p. e1002179, 2011.

DOLINSKY, T. J. et al. **PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations.** *Nucleic acids research*, v. 35, n. Web Server issue, p. W522–5, 2007.

DOSZTÁNYI, Z. et al. **IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content.** *Bioinformatics (Oxford, England)*, v. 21, n. 16, p. 3433–4, 2005.

EDDY, S. R. **Profile hidden Markov models**. *Bioinformatics*, v. 14, n. 9, p. 755–763, 1998.

ESWAR, N. et al. **Protein structure modeling with MODELLER.** *Methods in molecular biology (Clifton, N.J.)*, v. 426, p. 145–59, 2008.

FADRNÁ, E.; HLADECKOVÁ, K.; KOCA, J. **Long-range electrostatic interactions in molecular dynamics: an endothelin-1 case study.** *Journal of biomolecular structure & dynamics*, v. 23, n. 2, p. 151–62, 2005.

FERNANDES, T. V. A. **Desenvolvimento e Aplicação de Métodos Computacionais para Predição de Estrutura de Proteínas**. [s.l.] Universidade Federal do Rio de Janeiro, 2014.

FISCHER, E. **Einfluss der Configuration auf die Wirkung der Enzyme**. *Berichte der deutschen chemischen Gesellschaft*, v. 27, n. 3, p. 2985–2993, 1894.

FLETCHER, J. A.; RUBIN, B. P. **KIT mutations in GIST.** *Current opinion in genetics & development*, v. 17, n. 1, p. 3–7, 2007.

FLOUDAS, C. A. **Computational methods in protein structure prediction.** *Biotechnology and bioengineering*, v. 97, n. 2, p. 207–13, 2007.

FORSTER, M. J. **Molecular modelling in structural biology**. *Micron*, v. 33, n. 4, p. 365–384, 2002.

FRIESNER, R. A. et al. **Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy.** *Journal of medicinal chemistry*, v. 47, n. 7, p. 1739–49, 2004.

FRISHMAN, D.; ARGOS, P. **Knowledge-based protein secondary structure assignment.** *Proteins*, v. 23, n. 4, p. 566–79, 1995.

FROST, M. J. et al. **Juxtamembrane mutant V560GKit is more sensitive to Imatinib (STI571) compared with wild-type c-kit whereas the kinase domain mutant D816VKit is resistant.** *Molecular cancer therapeutics*, v. 1, n. 12, p. 1115–24, 2002.

GAJIWALA, K. S. et al. **KIT kinase mutants show unique mechanisms of drug resistance to imatinib and sunitinib in gastrointestinal stromal tumor patients.** *Proceedings of the National Academy of Sciences of the United States of America*, v. 106, n. 5, p. 1542–7, 2009.

GARCEAU, V. et al. **Pivotal Advance: Avian colony-stimulating factor 1 (CSF-1), interleukin-34 (IL-34), and CSF-1 receptor genes and gene products.** *Journal of leukocyte biology*, v. 87, n. 5, p. 753–64, 2010.

GARCÍA, A. **Large-amplitude nonlinear motions in proteins.** *Physical review letters*, v. 68, n. 17, p. 2696–2699, 1992.

GARNIER, J.; GIBRAT, J. F.; ROBSON, B. **GOR method for predicting protein secondary structure from amino acid sequence.** *Methods in enzymology*, v. 266, p. 540–53, 1996.

GEOURJON, C.; DELÉAGE, G. **SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments**. *Bioinformatics*, v. 11, n. 6, p. 681–684, 1995.

GLOVER, H. R. et al. **Selection of activating mutations of c-fms in FDC-P1 cells.** *Oncogene*, v. 11, n. 7, p. 1347–56, 1995.

GOLDBERG, H. J. et al. **The Tyrosine Kinase-Activity of the Epidermal-Growth-Factor Receptor Is Necessary for Phospholipase-A2 Activation**. *Biochemical Journal*, v. 267, n. 2, p. 461–465, 1990.

GORRE, M. E. et al. **Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification.** *Science (New York, N.Y.)*, v. 293, n. 5531, p. 876–80, 2001.

GOUNDER, M. M.; MAKI, R. G. **Molecular basis for primary and secondary tyrosine kinase inhibitor resistance in gastrointestinal stromal tumor.** *Cancer chemotherapy and pharmacology*, v. 67 Suppl 1, p. S25–43, 2011.

GRICHNIK, J. M. et al. **The SCF/KIT pathway plays a critical role in the control of normal human melanocyte homeostasis.** *The Journal of investigative dermatology*, v. 111, n. 2, p. 233–8, 1998.

GRIFFITH, J. et al. **The Structural Basis for Autoinhibition of FLT3 by the Juxtamembrane Domain**. *Molecular Cell*, v. 13, n. 2, p. 169–178, 2004.

GRONT, D. et al. **Generalized fragment picking in Rosetta: design, protocols and applications.** *PloS one*, v. 6, n. 8, p. e23294, 2011.

GU, J. et al. **MoDock: A multi-objective strategy improves the accuracy for molecular docking.** *Algorithms for molecular biology : AMB*, v. 10, p. 8, 2015.

HALGREN, T. A. et al. **Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening.** *Journal of medicinal chemistry*, v. 47, n. 7, p. 1750–9, 2004.

HALPERIN, I. et al. **Principles of docking: An overview of search algorithms and a guide to scoring functions**. *Proteins-Structure Function and Genetics*, v. 47, n. 4, p. 409–443, 2002.

HANDL, J. et al. **The dual role of fragments in fragment-assembly methods for de novo protein structure prediction.** *Proteins*, v. 80, n. 2, p. 490–504, 2012.

HARDIN, C.; POGORELOV, T. V; LUTHEY-SCHULTEN, Z. **Ab initio protein structure prediction.** *Current opinion in structural biology*, v. 12, n. 2, p. 176–81, 2002.

HARIR, N. et al. **Oncogenic Kit controls neoplastic mast cell growth through a Stat5/PI3-kinase signaling cascade.** *Blood*, v. 112, n. 6, p. 2463–73, 2008.

HAYASHI, K. et al. **Invasion activating caveolin-1 mutation in human scirrhous breast cancers**. *Cancer Research*, v. 61, n. 6, p. 2361–2364, 2001.

HAYWARD, S.; DE GROOT, B. L. **Normal modes and essential dynamics.** *Methods in molecular biology (Clifton, N.J.)*, v. 443, p. 89–106, 2008.

HEINRICH, M. C. et al. **Molecular correlates of imatinib resistance in gastrointestinal stromal tumors.** *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, v. 24, n. 29, p. 4764–74, 2006.

HESS, B. et al. **LINCS: A linear constraint solver for molecular simulations**. *Journal of Computational Chemistry*, v. 18, n. 12, p. 1463–1472, 1997.

HESS, B. **P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation**. *Journal of Chemical Theory and Computation*, v. 4, n. 1, p. 116–122, 2008.

HILLISCH, A.; PINEDA, L. F.; HILGENFELD, R. **Utility of homology models in the drug discovery process**. *Drug Discovery Today*, v. 9, n. 15, p. 659–669, 2004.

HOMEYER, N.; GOHLKE, H. **Free Energy Calculations by the Molecular Mechanics Poisson−Boltzmann Surface Area Method**. *Molecular Informatics*, v. 31, n. 2, p. 114–122, 2012.

HONEGGER, A. M. **Separate endocytic pathways of kinase-defective and -active EGF receptor mutants expressed in same cells**. *The Journal of Cell Biology*, v. 110, n. 5, p. 1541–1548, 1990.

HOOFT, R. W. W. et al. **Errors in protein structures**. *Nature*, v. 381, n. 6580, p. 272, 1996.

HORNAK, V. et al. **Comparison of multiple Amber force fields and development of improved protein backbone parameters.** *Proteins*, v. 65, n. 3, p. 712–25, 2006.

HUANG, S.-Y.; GRINTER, S. Z.; ZOU, X. **Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions.** *Physical chemistry chemical physics : PCCP*, v. 12, n. 40, p. 12899–908, 2010.

HUANG, S.-Y.; ZOU, X. **Advances and challenges in protein-ligand docking.** *International journal of molecular sciences*, v. 11, n. 8, p. 3016–34, 2010.

HUBBARD, S. R.; MILLER, W. T. **Receptor tyrosine kinases: mechanisms of activation and signaling.** *Current opinion in cell biology*, v. 19, n. 2, p. 117–23, 2007.

HUBBARD, S. R.; TILL, J. H. **Protein tyrosine kinase structure and function.** *Annual review of biochemistry*, v. 69, p. 373–98, 2000.

HUMPHREY, W.; DALKE, A.; SCHULTEN, K. **VMD: Visual molecular dynamics**. *Journal of Molecular Graphics*, v. 14, n. 1, p. 33–38, 1996.

HÜNENBERGER, P. H.; MCCAMMON, J. A. **Ewald artifacts in computer simulations of ionic solvation and ion–ion interaction: A continuum electrostatics study**. *The Journal of Chemical Physics*, v. 110, n. 4, p. 1856, 1999.

IHAKA, R.; GENTLEMAN, R. **R: A Language for Data Analysis and Graphics**. *Journal of Computational and Graphical Statistics*, v. 5, n. 3, p. 299–314, 1996.

JACOBSON, M. P. et al. **A hierarchical approach to all-atom protein loop prediction.** *Proteins*, v. 55, n. 2, p. 351–67, 2004.

JAMROZ, M.; KOLINSKI, A. **Modeling of loops in proteins: a multi-method approach.** *BMC structural biology*, v. 10, n. 1, p. 5, 2010.

JONES, D. T. **Protein secondary structure prediction based on position-specific scoring matrices.** *Journal of molecular biology*, v. 292, n. 2, p. 195–202, 1999.

JONES, D. T.; TAYLOR, W. R.; THORNTON, J. M. **A new approach to protein fold recognition.** *Nature*, v. 358, n. 6381, p. 86–9, 1992.

JOOS, H. et al. **Tyrosine phosphorylation of the juxtamembrane domain of the v-Fms oncogene product is required for its association with a 55-kDa protein.** *The Journal of biological chemistry*, v. 271, n. 40, p. 24476–81, 1996.

JORGENSEN, W. L.; JENSON, C. **Temperature dependence of TIP3P, SPC, and TIP4P water from NPT Monte Carlo simulations: Seeking temperatures of maximum density**. *Journal of Computational Chemistry*, v. 19, n. 10, p. 1179–1186, 1998.

KABSCH, W.; SANDER, C. **Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.** *Biopolymers*, v. 22, n. 12, p. 2577–637, 1983.

KAR, G. et al. **Allostery and population shift in drug discovery.** *Current opinion in pharmacology*, v. 10, n. 6, p. 715–22, 2010.

KARPLUS, K. **SAM-T08, HMM-based protein structure prediction.** *Nucleic acids research*, v. 37, n. Web Server issue, p. W492–7, 2009.

KARPLUS, M.; KUSHICK, J. N. **Method for Estimating the Configurational Entropy of Macromolecules**. *Macromolecules*, v. 14, n. 2, p. 325–332, 1981.

KERR, D. J. **K. N. Syrigos, K. Harrington (eds). Targeted Therapy for Cancer**. *Annals of Oncology* , v. 14 , n. 8 , p. 1333, 2003.

KIM, D. E.; CHIVIAN, D.; BAKER, D. **Protein structure prediction and analysis using the Robetta server.** *Nucleic acids research*, v. 32, n. Web Server issue, p. W526–31, 2004.

KITCHEN, D. B. et al. **Docking and scoring in virtual screening for drug discovery: Methods and applications**. *Nature Reviews Drug Discovery*, v. 3, n. 11, p. 935–949, 2004.

KOLLMAN, P. **Free energy calculations: Applications to chemical and biochemical phenomena**. *Chemical Reviews*, v. 93, p. 2395–2417, 1993.

KOLLMAN, P. A. et al. **Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models**. *Accounts of Chemical Research*, v. 33, n. 12, p. 889–897, 2000.

KOSHLAND, D. E. **Application of a Theory of Enzyme Specificity to Protein Synthesis**. *Proceedings of the National Academy of Sciences of the United States of America*, v. 44, n. 2, p. 98–104, 1958.

KOSHLAND, D. E.; NÉMETHY, G.; FILMER, D. **Comparison of experimental binding data and theoretical models in proteins containing subunits.** *Biochemistry*, v. 5, n. 1, p. 365–85, 1966.

KRAUTLER, V.; VAN GUNSTEREN, W. F.; HUNENBERGER, P. H. **A fast SHAKE: Algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations**. *Journal of Computational Chemistry*, v. 22, n. 5, p. 501–508, 2001.

KUHLMAN, B. et al. **Design of a novel globular protein fold with atomic-level accuracy.** *Science (New York, N.Y.)*, v. 302, n. 5649, p. 1364–8, 2003.

KUMARI, R. et al. **g_mmpbsa - A GROMACS tool for high-throughput MM-PBSA calculations.** *Journal of chemical information and modeling*, 2014.

KURIYAN, J.; COWBURN, D. **Modular peptide recognition domains in eukaryotic signaling.** *Annual review of biophysics and biomolecular structure*, v. 26, p. 259–88, 1997.

LAINE, E. et al. **Mutation D816V alters the internal structure and dynamics of c-KIT receptor cytoplasmic region: implications for dimerization and activation mechanisms.** *PLoS computational biology*, v. 7, n. 6, p. e1002068, 2011.

LAINE, E.; AUCLAIR, C.; TCHERTANOV, L. **Allosteric communication across the native and mutated KIT receptor tyrosine kinase.** *PLoS computational biology*, v. 8, n. 8, p. e1002661, 2012.

LANDAU, M.; BEN-TAL, N. **Dynamic equilibrium between multiple active and inactive conformations explains regulation and oncogenic mutations in ErbB receptors.** *Biochimica et biophysica acta*, v. 1785, n. 1, p. 12–31, 2008.

LASKER, K. et al. **Macromolecular assembly structures by comparative modeling and electron microscopy.** *Methods in molecular biology (Clifton, N.J.)*, v. 857, p. 331–50, 2012.

LASKOWSKI, R. A. et al. **Procheck - a Program to Check the Stereochemical Quality of Protein Structures**. *Journal of Applied Crystallography*, v. 26, p. 283–291, 1993.

LEACH, A. R. **Molecular Modelling: Principles and Applications**. illustrate ed.[s.l.] Prentice Hall, 2001.

LEAVER-FAY, A. et al. **ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules.** *Methods in enzymology*, v. 487, p. 545–74, 2011.

LEE, M. R.; DUAN, Y.; KOLLMAN, P. A. **Use of MM-PB/SA in estimating the free energies of proteins: Application to native, intermediates, and unfolded villin headpiece**. *Proteins-Structure Function and Genetics*, v. 39, n. 4, p. 309–316, 2000.

LEMMON, M. A.; SCHLESSINGER, J. **Cell signaling by receptor tyrosine kinases.** *Cell*, v. 141, n. 7, p. 1117–34, 2010.

LEVIN, J. M.; ROBSON, B.; GARNIER, J. **An algorithm for secondary structure determination in proteins based on sequence similarity**. *FEBS Letters*, v. 205, n. 2, p. 303–308, 1986.

LEVINTHAL, C. **How to fold graciously**. *Mossbaun Spectroscopy in Biological Systems Proceedings*, v. 67, n. 41, p. 22–24, 1969.

LEVITT, M.; SANDER, C.; STERN, P. S. **Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme.** *Journal of molecular biology*, v. 181, n. 3, p. 423–47, 1985.

LI, J.; HAN, W.; YU, Y. **Protein Engineering - Technology and Application**. [s.l.] InTech, 2013.

LI, S. et al. **Structural and biochemical evidence for an autoinhibitory role for tyrosine 984 in the juxtamembrane region of the insulin receptor.** *The Journal of biological chemistry*, v. 278, n. 28, p. 26007–14, 2003.

LICHTENTHALER, F. W. **100 Years"Schlüssel-Schloss-Prinzip": What Made Emil Fischer Use this Analogy?** *Angewandte Chemie International Edition in English*, v. 33, n. 2324, p. 2364–2374, 1995.

LIN, H. et al. **Discovery of a cytokine and its receptor by functional screening of the extracellular proteome.** *Science (New York, N.Y.)*, v. 320, n. 5877, p. 807–11, 2008.

LIU, T.; TANG, G. W.; CAPRIOTTI, E. **Comparative modeling: the state of the art and protein drug target structure prediction.** *Combinatorial chemistry & high throughput screening*, v. 14, n. 6, p. 532–47, 2011.

LIU, Y.; GRAY, N. S. **Rational design of inhibitors that bind to inactive kinase conformations.** *Nature chemical biology*, v. 2, n. 7, p. 358–64, 2006.

LÜTHY, R.; BOWIE, J. U.; EISENBERG, D. **Assessment of protein models with three-dimensional profiles.** *Nature*, v. 356, n. 6364, p. 83–5, 1992.

LYMAN, E.; ZUCKERMAN, D. M. **Ensemble-based convergence analysis of biomolecular trajectories.** *Biophysical journal*, v. 91, n. 1, p. 164–72, 2006.

MA, J.; KARPLUS, M. **The allosteric mechanism of the chaperonin GroEL: A dynamic analysis**. *Proceedings of the National Academy of Sciences*, v. 95, n. 15, p. 8502–8507, 1998.

MACKERELL, A. D. et al. **All-atom empirical potential for molecular modeling and dynamics studies of proteins.** *The journal of physical chemistry. B*, v. 102, n. 18, p. 3586–616, 1998.

MACKERELL, A. D.; BANAVALI, N.; FOLOPPE, N. **Development and current status of the CHARMM force field for nucleic acids.** *Biopolymers*, v. 56, n. 4, p. 257–65, 2000.

MACKERELL, A. D.; FEIG, M.; BROOKS, C. L. **Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations.** *Journal of computational chemistry*, v. 25, n. 11, p. 1400–15, 2004.

MANDELL, D. J.; COUTSIAS, E. A.; KORTEMME, T. **Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling.** *Nature methods*, v. 6, n. 8, p. 551–2, 2009.

MARTÍ-RENOM, M. A. et al. **Comparative protein structure modeling of genes and genomes.** *Annual review of biophysics and biomolecular structure*, v. 29, p. 291–325, 2000.

MATHIEU BASTIAN, S. H. **Gephi: An Open Source Software for Exploring and Manipulating Networks.**2009

MCCULLOCH, E. A.; MINDEN, M. D. **The cell surface receptor encoded by the proto-oncogene KIT and its ligand.** *Cancer treatment and research*, v. 64, p. 45–77, 1993.

MCDERMOTT, R. S. et al. **Circulating macrophage colony stimulating factor as a marker of tumour progression**. *European Cytokine Network*, v. 13, n. 1, p. 121–7, 2002.

MCDONALD, I. K.; THORNTON, J. M. **Satisfying hydrogen bonding potential in proteins.** *Journal of molecular biology*, v. 238, n. 5, p. 777–93, 1994.

MCGUFFIN, L. J.; BRYSON, K.; JONES, D. T. **The PSIPRED protein structure prediction server**. *Bioinformatics*, v. 16, n. 4, p. 404–405, 2000.

MCWILLIAM, H. et al. **Analysis Tool Web Services from the EMBL-EBI.** *Nucleic acids research*, v. 41, n. Web Server issue, p. W597–600, 2013.

MEIROVITCH, H. **Recent developments in methodologies for calculating the entropy and free energy of biological systems by computer simulation.** *Current opinion in structural biology*, v. 17, n. 2, p. 181–6, 2007.

MENKE, J. et al. **Circulating CSF-1 promotes monocyte and macrophage phenotypes that enhance lupus nephritis.** *Journal of the American Society of Nephrology : JASN*, v. 20, n. 12, p. 2581–92, 2009.

MEYERS, M. J. et al. **Structure-based drug design enables conversion of a DFG-in binding CSF-1R kinase inhibitor to a DFG-out binding mode.** *Bioorganic & medicinal chemistry letters*, v. 20, n. 5, p. 1543–7, 2010.

MILLER, B. R. et al. **MMPBSA.py : An Efficient Program for End-State Free Energy Calculations**. *Journal of Chemical Theory and Computation*, v. 8, n. 9, p. 3314–3321, 2012.

MILLER, M. D. et al. **FLOG: a system to select "quasi-flexible" ligands complementary to a receptor of known three-dimensional structure.** *Journal of computer-aided molecular design*, v. 8, n. 2, p. 153–74, 1994.

MIYAMOTO, S.; KOLLMAN, P. A. **Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models**. *Journal of Computational Chemistry*, v. 13, n. 8, p. 952–962, 1992.

MOL, C. D. et al. **Structure of a c-kit product complex reveals the basis for kinase transactivation.** *The Journal of biological chemistry*, v. 278, n. 34, p. 31461–4, 2003.

MOL, C. D. et al. **Structural basis for the autoinhibition and STI-571 inhibition of c-Kit tyrosine kinase.** *The Journal of biological chemistry*, v. 279, n. 30, p. 31655–63, 2004.

MONOD, J.; WYMAN, J.; CHANGEUX, J.-P. **On the nature of allosteric transitions: A plausible model**. *Journal of Molecular Biology*, v. 12, n. 1, p. 88–118, 1965.

MORLEY, G. M. et al. **Cell specific transformation by c-fms activating loop mutations is attributable to constitutive receptor degradation.** *Oncogene*, v. 18, n. 20, p. 3076–84, 1999.

MORRIS, G. M. et al. **Distributed automated docking of flexible ligands to proteins: Parallel applications of AutoDock 2.4**. *Journal of Computer-Aided Molecular Design*, v. 10, n. 4, p. 293–304, 1996.

MOTLAGH, H. N. et al. **The ensemble nature of allostery.** *Nature*, v. 508, n. 7496, p. 331–9, 2014.

MOUCHEMORE, K. A.; PIXLEY, F. J. **CSF-1 signaling in macrophages: pleiotrophy through phosphotyrosine-based signaling pathways.** *Critical reviews in clinical laboratory sciences*, v. 49, n. 2, p. 49–61, 2012.

MOULT, J. et al. **Critical assessment of methods of protein structure prediction (CASP)--round x.** *Proteins*, v. 82 Suppl 2, p. 1–6, 2014.

NAGAR, B. **c-Abl Tyrosine Kinase and Inhibition by the Cancer Drug Imatinib (Gleevec/STI-571)**. *J. Nutr.*, v. 137, n. 6, p. 1518S–1523, 2007.

NAYEEM, A.; SITKOFF, D.; KRYSTEK, S. **A comparative study of available software for high-accuracy homology modeling: From sequence alignments to structural models**. *Protein Science*, v. 15, n. 4, p. 808–824, 2006.

NING, Z. Q.; LI, J.; ARCECI, R. J. **Signal transducer and activator of transcription 3 activation is required for Asp(816) mutant c-Kit-mediated cytokine-independent survival and proliferation in human leukemia cells.** *Blood*, v. 97, n. 11, p. 3559–67, 2001.

NYBERG, A. M.; SCHLICK, T. **Increasing the time step in molecular dynamics**. *Chemical Physics Letters*, v. 198, n. 6, p. 538–546, 1992.

ONUFRIEV, A.; BASHFORD, D.; CASE, D. A. **Modification of the Generalized Born Model Suitable for Macromolecules**. *The Journal of Physical Chemistry B*, v. 104, n. 15, p. 3712–3720, 2000.

ONUFRIEV, A.; BASHFORD, D.; CASE, D. A. **Exploring protein native states and large-scale conformational changes with a modified generalized born model.** *Proteins*, v. 55, n. 2, p. 383–94, 2004.

PANDINI, A. et al. **Detection of allosteric signal transmission by information-theoretic analysis of protein dynamics.** *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, v. 26, n. 2, p. 868–81, 2012.

PARDANANI, A. **Systemic mastocytosis in adults: 2013 update on diagnosis, risk stratification, and management.** *American journal of hematology*, v. 88, n. 7, p. 612–24, 2013.

PARENTI, M. D.; RASTELLI, G. **Advances and applications of binding affinity prediction methods in drug discovery.** *Biotechnology advances*, v. 30, n. 1, p. 244–50, 2012.

PATSIALOU, A. et al. **Invasion of human breast cancer cells in vivo requires both paracrine and autocrine loops involving the colony-stimulating factor-1 receptor.** *Cancer research*, v. 69, n. 24, p. 9498–506, 2009.

PENEV, P.; ATICK, J. **Local feature analysis: a general statistical theory for object representation**. *Network: Computation in Neural Systems*, v. 7, n. 3, p. 477–500, 1996.

PERAHIA, D.; MOUAWAD, L. **Computation of low-frequency normal modes in macromolecules: Improvements to the method of diagonalization in a mixed basis and application to hemoglobin**. *Computers & Chemistry*, v. 19, n. 3, p. 241–246, 1995.

PERUTZ, M. F. **Stereochemistry of cooperative effects in haemoglobin.** *Nature*, v. 228, n. 5273, p. 726–39, 1970.

PERUTZ, M. F. et al. **The stereochemical mechanism of the cooperative effects in hemoglobin revisited.** *Annual review of biophysics and biomolecular structure*, v. 27, p. 1–34, 1998.

PETERSEN, B. et al. **A generic method for assignment of reliability scores applied to solvent accessibility predictions.** *BMC structural biology*, v. 9, n. 1, p. 51, 2009.

PIANA, S.; KLEPEIS, J. L.; SHAW, D. E. **Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations.** *Current opinion in structural biology*, v. 24, p. 98–105, 2014.

PIAZZA, F.; SANEJOUAND, Y.-H. **Long-range energy transfer in proteins.** *Physical biology*, v. 6, n. 4, p. 046014, 2009.

PIXLEY, F. J.; STANLEY, E. R. **CSF-1 regulation of the wandering macrophage: complexity in action**. *Trends in Cell Biology*, v. 14, n. 11, p. 628–638, 2004.

POLLASTRI, G. et al. **Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles.** *Proteins*, v. 47, n. 2, p. 228–35, 2002.

POLLASTRI, G.; MCLYSAGHT, A. **Porter: a new, accurate server for protein secondary structure prediction.** *Bioinformatics (Oxford, England)*, v. 21, n. 8, p. 1719–20, 2005.

PRICE, C. J.; GREEN, J.; KIRSNER, R. S. **Mastocytosis in Children Is Associated with Mutations in c-KIT**. *Journal of Investigative Dermatology*, v. 130, n. 3, p. 639, 2010.

PYONTECK, S. M. et al. **CSF-1R inhibition alters macrophage polarization and blocks glioma progression.** *Nature medicine*, v. 19, n. 10, p. 1264–72, 2013.

RAMACHANDRAN, G. N.; RAMAKRISHNAN, C.; SASISEKHARAN, V. **Stereochemistry of Polypeptide Chain Configurations**. *Journal of Molecular Biology*, v. 7, n. 1, p. 95–&, 1963.

RAREY, M. et al. **A fast flexible docking method using an incremental construction algorithm**. *Journal of Molecular Biology*, v. 261, n. 3, p. 470–489, 1996.

RASTELLI, G. et al. **Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA.** *Journal of computational chemistry*, v. 31, n. 4, p. 797–810, 2010.

RICHARDSEN, E. et al. **The prognostic impact of M-CSF, CSF-1 receptor, CD68 and CD3 in prostatic carcinoma.** *Histopathology*, v. 53, n. 1, p. 30–8, 2008.

RIDGE, S. A. et al. **FMS mutations in myelodysplastic, leukemic, and normal subjects.** *Proceedings of the National Academy of Sciences of the United States of America*, v. 87, n. 4, p. 1377–80, 1990.

ROBINSON, D. R.; WU, Y. M.; LIN, S. F. **The protein tyrosine kinase family of the human genome.** *Oncogene*, v. 19, n. 49, p. 5548–57, 2000.

ROHL, C. A. et al. **Protein structure prediction using Rosetta.** *Methods in enzymology*, v. 383, p. 66–93, 2004.

ROSNET, O. et al. **Human Flt3/Flk2 Gene - Cdna Cloning and Expression in Hematopoietic-Cells**. *Blood*, v. 82, n. 4, p. 1110–1119, 1993.

ROSNET, O.; BIRNBAUM, D. **Hematopoietic Receptors of Class-Iii Receptor-Type Tyrosine Kinases**. *Critical Reviews in Oncogenesis*, v. 4, n. 6, p. 595–613, 1993.

ROST, B. **Twilight zone of protein sequence alignments**. *Protein Engineering Design and Selection*, v. 12, n. 2, p. 85–94, 1999.

ROST, B. **Review: Protein Secondary Structure Prediction Continues to Rise**. *Journal of Structural Biology*, v. 134, n. 2-3, p. 204–218, 2001.

SALI, A.; OVERINGTON, J. P. **Derivation of rules for comparative protein modeling from a database of protein structure alignments.** *Protein science : a publication of the Protein Society*, v. 3, n. 9, p. 1582–96, 1994.

SANNER, M. F.; OLSON, A. J.; SPEHNER, J. C. **Reduced surface: an efficient way to compute molecular surfaces.** *Biopolymers*, v. 38, n. 3, p. 305–20, 1996.

SCHINDLER, T. et al. **Structural mechanism for STI-571 inhibition of abelson tyrosine kinase.** *Science (New York, N.Y.)*, v. 289, n. 5486, p. 1938–42, 2000.

SCHLESSINGER, J.; LEMMON, M. A. **SH2 and PTB domains in tyrosine kinase signaling.** *Science's STKE : signal transduction knowledge environment*, v. 2003, n. 191, p. RE12, 2003.

SCHLICK, T. **Molecular Modeling and Simulation**. New York, NY: Springer New York, 2002. v. 21

SCHOLL, S. M. et al. **Circulating levels of colony-stimulating factor 1 as a prognostic indicator in 82 patients with epithelial ovarian cancer.** *British journal of cancer*, v. 69, n. 2, p. 342–6, 1994.

Schrödinger Release 2014-2: LigPrep. , 2014.

Schrödinger Release 2014-2: Maestro. , 2014.

SCHWEDE, T. **Protein modeling: what happened to the "protein structure gap"?** *Structure (London, England : 1993)*, v. 21, n. 9, p. 1531–40, 2013.

SEELIGER, M. A. et al. **c-Src binds to the cancer drug imatinib with an inactive Abl/c-Kit conformation and a distributed thermodynamic penalty.** *Structure (London, England : 1993)*, v. 15, n. 3, p. 299–311, 2007.

SHAW, D. E. et al. **Anton, a special-purpose machine for molecular dynamics simulation**Proceedings of the 34th annual international symposium on Computer architecture - ISCA '07. **Anais**...New York, New York, USA: ACM Press, 2007Disponível em: <http://dl.acm.org/citation.cfm?id=1250662.1250664>. Acesso em: 19 mar. 2015

SHEN, M.-Y.; SALI, A. **Statistical potential for assessment and prediction of protein structures.** *Protein science : a publication of the Protein Society*, v. 15, n. 11, p. 2507–24, 2006.

SHERMAN, W. et al. **Novel procedure for modeling ligand/receptor induced fit effects.** *Journal of medicinal chemistry*, v. 49, n. 2, p. 534–53, 2006.

SKJAERVEN, L.; MARTINEZ, A.; REUTER, N. **Principal component and normal mode analysis of proteins; a quantitative comparison using the GroEL subunit.** *Proteins*, v. 79, n. 1, p. 232–43, 2011.

Small-Molecule Drug Discovery Suite: Induced Fit Docking protocol. , 2014.

SOARES, M. J. et al. **CSF1R copy number changes, point mutations, and RNA and protein overexpression in renal cell carcinomas.** *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*, v. 22, n. 6, p. 744–52, 2009.

SUCH, E. et al. **Absence of mutations in the tyrosine kinase and juxtamembrane domains of C-FMS gene in chronic myelomonocytic leukemia (CMML).** *Leukemia research*, v. 33, n. 9, p. e162–3, 2009.

SÜEL, G. M. et al. **Evolutionarily conserved networks of residues mediate allosteric communication in proteins.** *Nature structural biology*, v. 10, n. 1, p. 59–69, 2003.

SUZEK, B. E. et al. **UniRef: comprehensive and non-redundant UniProt reference clusters.** *Bioinformatics (Oxford, England)*, v. 23, n. 10, p. 1282–8, 2007.

SWOPE, W. C. **A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters**. *The Journal of Chemical Physics*, v. 76, n. 1, p. 637, 1982.

SWORDS, R.; FREEMAN, C.; GILES, F. **Targeting the FMS-like tyrosine kinase 3 in acute myeloid leukemia.** *Leukemia*, v. 26, n. 10, p. 2176–85, 2012.

SZAKÁCS, Z. et al. **Acid-base profiling of imatinib (gleevec) and its fragments.** *Journal of medicinal chemistry*, v. 48, n. 1, p. 249–55, 2005.

TAKESHITA, S. et al. **c-Fms Tyrosine 559 Is a Major Mediator of M-CSF-induced Proliferation of Primary Macrophages**. *Journal of Biological Chemistry*, v. 282, n. 26, p. 18980–18990, 2007.

TAMA, F.; SANEJOUAND, Y.-H. **Conformational change of proteins arising from normal mode calculations**. *Protein Engineering Design and Selection*, v. 14, n. 1, p. 1–6, 2001.

TAMBORINI, E. et al. **A new mutation in the KIT ATP pocket causes acquired resistance to imatinib in a gastrointestinal stromal tumor patient**. *Gastroenterology*, v. 127, n. 1, p. 294–299, 2004.

TAYLOR, J. R. et al. **FMS receptor for M-CSF (CSF-1) is sensitive to the kinase inhibitor imatinib and mutation of Asp-802 to Val confers resistance.** *Oncogene*, v. 25, n. 1, p. 147–51, 2006.

TOPF, M.; SALI, A. **Combining electron microscopy and comparative protein structure modeling.** *Current opinion in structural biology*, v. 15, n. 5, p. 578–85, 2005.

TOPHAM, C. M. et al. **Comparative modelling of major house dust mite allergen Der p I: structure validation using an extended environmental amino acid propensity table.** *Protein engineering*, v. 7, n. 7, p. 869–94, 1994.

TSAI, C.-J.; DEL SOL, A.; NUSSINOV, R. **Allostery: absence of a change in shape does not imply that allostery is not at play.** *Journal of molecular biology*, v. 378, n. 1, p. 1–11, 2008.

ULLRICH, A.; SCHLESSINGER, J. **Signal transduction by receptors with tyrosine kinase activity**. *Cell*, v. 61, n. 2, p. 203–212, 1990.

VAN DER SPOEL, D. et al. **GROMACS: fast, flexible, and free.** *Journal of computational chemistry*, v. 26, n. 16, p. 1701–18, 2005.

VAN GUNSTEREN, W. F.; BERENDSEN, H. J. C. **A Leap-frog Algorithm for Stochastic Dynamics**. *Molecular Simulation*, v. 1, n. 3, p. 173–185, 1988a.

VAN GUNSTEREN, W. F.; BERENDSEN, H. J. C. **A Leap-frog Algorithm for Stochastic Dynamics**. *Molecular Simulation*, v. 1, n. 3, p. 173–185, 1988b.

VANAALTEN, D. M. F. et al. **A comparison of techniques for calculating protein essential dynamics**. *Journal of Computational Chemistry*, v. 18, n. 2, p. 169–181, 1997.

VANGUNSTEREN, W. F.; BERENDSEN, H. J. C. **Computer-Simulation of Molecular-Dynamics - Methodology, Applications, and Perspectives in Chemistry**. *Angewandte Chemie-International Edition in English*, v. 29, n. 9, p. 992–1023, 1990.

VENKATACHALAM, C. M. et al. **LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites.** *Journal of molecular graphics & modelling*, v. 21, n. 4, p. 289–307, 2003.

VERDONK, M. L. et al. **Improved protein-ligand docking using GOLD.** *Proteins*, v. 52, n. 4, p. 609–23, 2003.

VERLET, L. **Computer Experiments on Classical Fluids .I. Thermodynamical Properties of Lennard-Jones Molecules**. *Physical Review*, v. 159, n. 1, p. 98–&, 1967.

VERSTRAETE, K.; SAVVIDES, S. N. **Extracellular assembly and activation principles of oncogenic class III receptor tyrosine kinases.** *Nature reviews. Cancer*, v. 12, n. 11, p. 753–66, 2012.

VILLARREAL, M. A.; MONTICH, G. G. **On the Ewald artifacts in computer simulations. The test-case of the octaalanine peptide with charged termini.** *Journal of biomolecular structure & dynamics*, v. 23, n. 2, p. 135–42, 2005.

VITA, M. et al. **Characterization of S628N: A Novel KIT Mutation Found in a Metastatic Melanoma.** *JAMA dermatology*, v. 150, n. 12, p. 1345–1349, 2014.

WALLACE, A. C.; LASKOWSKI, R. A.; THORNTON, J. M. **LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions**. *"Protein Engineering, Design and Selection",* v. 8, n. 2, p. 127–134, 1995.

WALTER, M. et al. **The 2.7 A crystal structure of the autoinhibited human c-Fms kinase domain.** *Journal of molecular biology*, v. 367, n. 3, p. 839–47, 2007.

WAN, S.; COVENEY, P. V. **Molecular dynamics simulation reveals structural and thermodynamic features of kinase activation by cancer mutations within the epidermal growth factor receptor.** *Journal of computational chemistry*, v. 32, n. 13, p. 2843–52, 2011.

WEBSTER, J. A. et al. **Variations in stromal signatures in breast and colorectal cancer metastases.** *The Journal of pathology*, v. 222, n. 2, p. 158–65, 2010.

WEI, S. et al. **Functional overlap but differential expression of CSF-1 and IL-34 in their CSF-1 receptor-mediated regulation of myeloid cells.** *Journal of leukocyte biology*, v. 88, n. 3, p. 495–505, 2010.

WIEDERSTEIN, M.; SIPPL, M. J. **ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins.** *Nucleic acids research*, v. 35, n. Web Server issue, p. W407–10, 2007.

XU, D.; WILLIAMSON, M. J.; WALKER, R. C. **Chapter 1 – Advancements in Molecular Dynamics Simulations of Biomolecules on Graphical Processing Units**. In: *Annual Reports in Computational Chemistry*. [s.l: s.n.]. v. 6p. 2–19.

YANG, L.-W. et al. **Principal component analysis of native ensembles of biomolecular structures (PCA_NEST): insights into functional dynamics.** *Bioinformatics (Oxford, England)*, v. 25, n. 5, p. 606–14, 2009.

YANG, P.-T. et al. **Increased expression of macrophage colony-stimulating factor in ankylosing spondylitis and rheumatoid arthritis.** *Annals of the rheumatic diseases*, v. 65, n. 12, p. 1671–2, 2006.

YANG, T. et al. **Virtual screening using molecular simulations.** *Proteins*, v. 79, n. 6, p. 1940–51, 2011.

YTREBERG, F. M.; SWENDSEN, R. H.; ZUCKERMAN, D. M. **Comparison of free energy methods for molecular systems.** *The Journal of chemical physics*, v. 125, n. 18, p. 184114, 2006.

YU, W. et al. **CSF-1 receptor structure/function in MacCsf1r-/- macrophages: regulation of proliferation, differentiation, and morphology.** *Journal of leukocyte biology*, v. 84, n. 3, p. 852–63, 2008.

YU, W. et al. **Macrophage proliferation is regulated through CSF-1 receptor tyrosines 544, 559, and 807.** *The Journal of biological chemistry*, v. 287, n. 17, p. 13694–704, 2012.

ZHANG, J.; YANG, P. L.; GRAY, N. S. **Targeting cancer with small molecule kinase inhibitors.** *Nature reviews. Cancer*, v. 9, n. 1, p. 28–39, 2009.

ZHANG, Z.; WRIGGERS, W. **Local feature analysis: a statistical theory for reproducible essential dynamics of large macromolecules.** *Proteins*, v. 64, n. 2, p. 391–403, 2006.

ZHAO, S.; IYENGAR, R. **Systems pharmacology: network analysis to identify multiscale mechanisms of drug action.** *Annual review of pharmacology and toxicology*, v. 52, p. 505–21, 2012.

ZOETE, V. et al. **SwissParam: a fast force field generation tool for small organic molecules.** *Journal of computational chemistry*, v. 32, n. 11, p. 2359–68, 2011.

ZOU, J. et al. **Detailed conformational dynamics of juxtamembrane region and activation loop in c-Kit kinase activation process.** *Proteins*, v. 72, n. 1, p. 323–32, 2008.

ZSEBO, K. M. et al. **Stem cell factor is encoded at the Sl locus of the mouse and is the ligand for the c-kit tyrosine kinase receptor.** *Cell*, v. 63, n. 1, p. 213–24, 1990.

ZUCCOTTO, F. et al. **Through the "gatekeeper door": exploiting the active kinase conformation.** *Journal of medicinal chemistry*, v. 53, n. 7, p. 2681–94, 2010.