جامعة بيرزيت
BIRZEIT UNIVERSITY

Identification of Druggable Binding sites in Ras Proteins using TACTICS

تحديد مواقع ربط الأدوية على سطح بروتينات راس باستخدام خوارزميه للتعلم الآلي

Rana Moqady

Supervisor: Prof. Abdallah Sayyed-Ahmad

February 2022

Identification of Druggable Binding sites in Ras Proteins using TACTICS

تحديد مواقع ربط الأدوية على سطح بروتينات راس باستخدام خوارزميه للتعلم الآلي

Rana Moqady

Supervisor: Prof. Abdallah Sayyed-Ahmad

This thesis was submitted in partial fulfillment of the requirements for the Master's Degree in Physics from the Faculty of Graduate Studies at Birzeit University, Palestine

February 2022

Identification of Druggable Binding sites in Ras Proteins using TACTICS

تحديد مواقع ربط الأدوية على سطح بروتينات راس باستخدام خوارزميه للتعلم الآلي

Rana Moqady

Accepted by the Faculty of Graduate Studies, Birzeit University, in partial fulfillment of the degree of Master of Physics

Thesis committee:

Abdallah Sayyed-Ahmad, Ph.D.

(Principal Advisor)

Wafaa Khater, Ph.D.

(Member)

Wael Karain, Ph.D.

(Member)

February 2022

I would like to dedicate this thesis to the pure soul of my father.

Acknowledgements

I would like to express my sincere gratitude to my advisor Prof. Abdallah Sayyed-Ahmad for the continuous support during this study. I would also like to thank him for his patience, providing motivation and sharing his immense knowledge. His guidance helped me to complete this research project and writing of this thesis. I could not have imagined having a better advisor and mentor for my master's study.

Finally, I would like express my very profound gratitude to my loving family for providing me with unfailing support and continuous encouragement throughout my years of study and throughout the process of writing this thesis. This accomplishment would not have been possible without them. Thank you.

Abstract

RAS proteins belong to the small GTPase family, which is classified as an allosteric enzyme that controls cell proliferation, differentiation, and development. As a result of critical Ras function, Mutations of Ras proteins are associated with $(15-20)\%$ of all human tumors. Several algorithms have been deployed to identify druggable binding sites on the Ras surface. In this work, we used a novel machine learning algorithm developed to identify druggable binding sites by analyzing MD simulations called TACTICS. Using TACTICS, we were able to robustly identify the known four cryptic Ras binding sites and investigate their presence in three K-Ras Mutants: G12D, G12DSOH/G12D, and pT32/G12D. Our results suggest that pocket p3 (residues: 107-111, 137,139) exists in all systems with a prominent appearance in pT32/G12D, where it occurs in the most different conformations. The appearance of p3 in three systems indicates that it is a druggable pocket. Pocket p4 (residues: 17, 21, 29, 31, 35) was most prominent in the G12D, while pocket p2 (residues: 61-65,92,99) in the oxidized G12DSOH/G12D counterpart.

ملخص

تنتمي بروتينات Ras إلى عائلة GTPase الصغيرة، تلعب بروتينات Ras دورا رئيسيا في التحكم في نمو وانقسام وتكاثر وتنظيم الموت المبرمج للخلايا من خلال نقل الإشارات داخل الخلية. لذلك حدوث خلل في وظيفة هذه البروتينات يؤدي إلى حدوث انواع مختلفه من الأورام السرطانية. تم في الأونة الأخيرة تطوير العديد من الخوارزميات من أجل تحديد مواقع ربط الأدوية على سطح البروتينات، والتي تم الكشف من خلالها عن أربعة مواقع ربط للأدوية خفية على سطح البروتين.

نستخدم في هذه الدراسة تقنية جديدة للبحث عن مواقع ربط الأدوية تدعى TACTICS ، وهي طريقة قائمة على دمج التعلم الآلي مع خوارزمية معقدة، لإيجاد مواقع الربط بدقة عالية. تم تطبيق هذه التقنية على ثلاث طفرات مختلفة لإظهار مواقع ربط الأدويةالموجودة في كل منها، هذه الدراسه مفيدة في المساعدة على تصميم الأدوية من خلال التركيز على مواقع ربط الأدوية الأكثر وضوحا ومن ثم البحث عن مركبات دوائية جديدة مناسبه لها.

Contents

List of Figures

List of Tables

Chapter 1: Introduction

Cancers are devastating diseases that affect the lives of millions of people around the world. They are caused by an uncontrolled growth of abnormal cells and dynamic alterations in the genome (which causes precancerous features in normal cells). The development of cancer impairs the normal biological process of healthy cells, invading nearby tissues and spreading them to distant tissues[1, 2]. In multicellular organisms, cellular behavior is controlled by a complex network of signaling pathways that regulates cell proliferation differentiation and growth. Mutations in genes that modulate these signaling pathways are thought to be one of the main causes of cancers[3, 4].

Ras small GTPase family of genes is frequently mutated in approximately 30% of all human tumors [5]. It was discovered in the 1960s as a viral component that caused the formation of sarcomas in rats [2, 4]. These proteins act as binary molecular switches controlling cell signaling pathways involved in cell division, proliferation, and development. There are three human Ras proteins: N-Ras, H-Ras, and K-Ras. Thyroid and bladder cancers are caused by mutations in H-Ras; Colorectal, 35 % lung, and 95 % pancreatic cancers are caused by mutations in K-Ras; 15% of melanoma is caused by mutations in N-Ras. Overall, K-Ras mutations in amino acids G12, G13, Q61 are observed in 85% of Ras-related cancers [4, 6–9]. Here, we are interested in K-Ras which consists of a catalytic domain (residues 1-

166), a membrane-binding region (residues 167 − 189), and a farnesyl anchor. The conserved catalytic domain includes six beta-strands, which make up the protein core, and is surrounded by five alpha-helices. The membrane binding region is not conserved and has notable sequence differences between Ras isoforms.

The catalytic domain of Ras interacts with effectors and exchange factors by modulating the conformations of two flexible canonical switches: switch 1 (SI: residues 30 − 38) and switch 2 (SII: residues 60 − 76) (see Figure 1.1). When Ras is bound to the nucleotide guanosine diphosphate (GDP), it becomes in the inactive "off state", and when it is bound to the nucleotide guanosine triphosphate (GTP), it becomes in the active "on state". In the active state Ras binds and activate its effector proteins such as RAF Kinesis, P13K, and RaLGDs. Ras enhanced activity by GTPase-activating proteins (GAPs) accelerates Ras inactivation. The exchange of the bound nucleotide GDP into GTP is facilitated by guanine nucleotide exchange factors (GEFs)[5, 10–13]. Therefore, developing K-Ras inhibitors is crucial to treat these fatal cancers.

Structure-based drug design is a complex process that includes many steps[14]. In this thesis, we focus on finding druggable cryptic binding sites on the surface of proteins. There are many methods used to identify drug binding sites. For example, FTMAP uses an energy mapping algorithm and a fast Fourier transform (FFT) correlation approach to identify druggable binding sites as "hot spots" by taking one PDB file as input. FTMAP searches the protein surface for regions that bind to a subset of 16 small organic probe molecules. By incorporating detailed energy expressions on grids, FTMAP allows us to locate low-energy probe clusters. Overlapping clusters of different probes are called consensus sites (CSs). FTMAP was applied to different proteins and yielded ligand binding sites found from NMR and X-ray crystallography techniques [15–18].

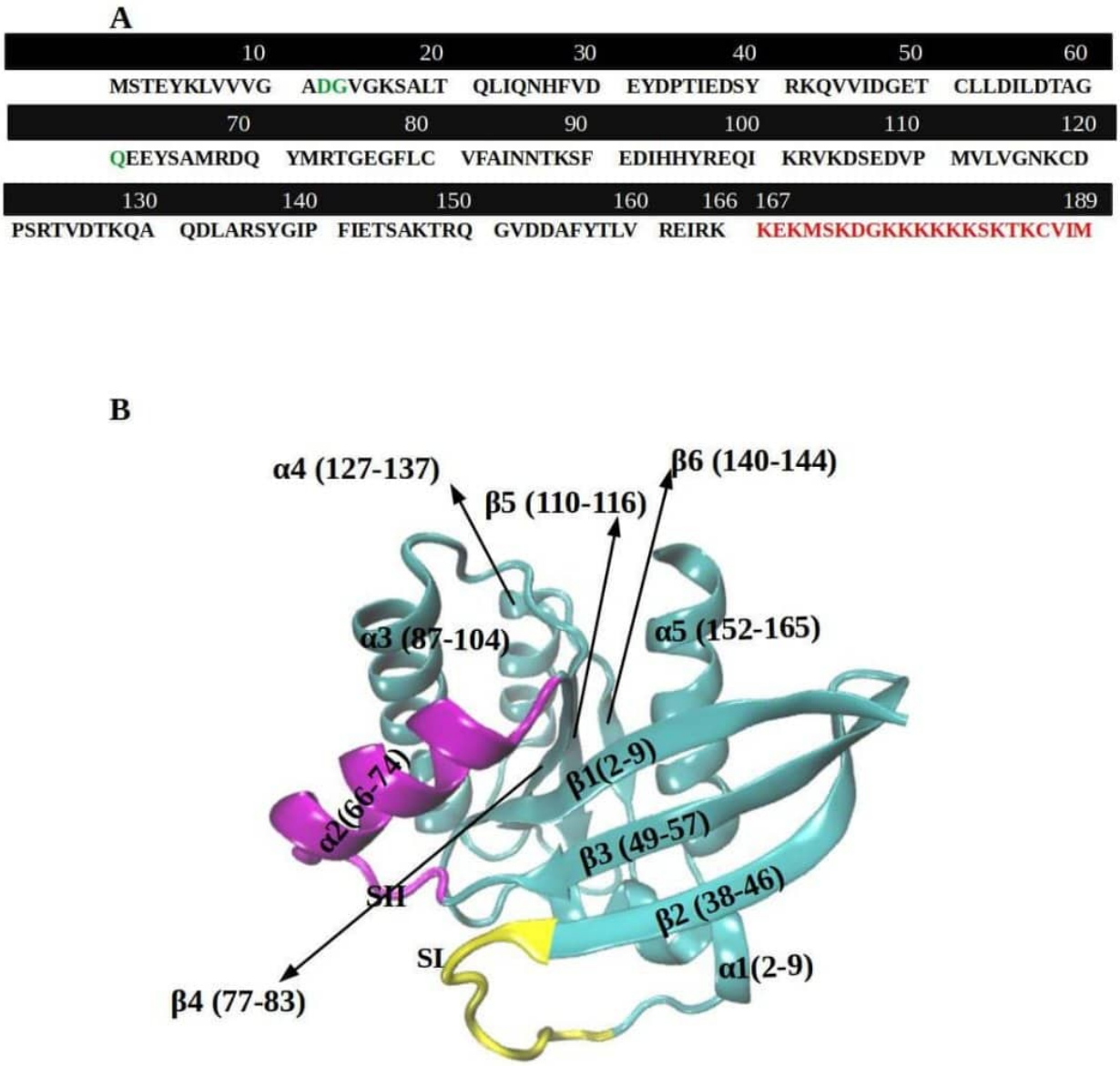

Figure 1.1: The sequence and structure of the G12D K-Ras. (A) The sequence of the amino acids of the catalytic domain (amino acids 1-166, three mutation sites are highlighted in green) and HVR (amino acids 167-189, are highlighted in red). (B) The structure is shown in cartoon representation. Switches SI and SII are colored yellow and magenta, respectively.

To identify druggable binding sites on K-RAS, Grant et al. [19] used FTMAP along with AutoLigand and BlindDock. AutoLigand determines the optimal ligand volume, shape, and atom type depending on the receptor properties. Therefore, regions of high affinity can be linked through low-affinity regions with the contiguous envelope technique. This should work as long as the affinity of the

volume is optimized [20]. A blind docking was performed using small molecules known to bind Ras. Essentially, this approach states that ligands will target a specific site if it offers an energetic or steric advantage over any other site [21]. These methods have been used to find four previously discovered sites as well as new sites. Three pockets, displaying low and high occupancy, were identified by FMAP (p1,p2,p3), and a new pocket called p3b was identified by AutoLigand. p3b is adjacent to p3 but is shallower and smaller than it. Hence it cannot be found by FTMAP or blind docking. They also discovered p3 and p4 through blind docking and found that p3 is the most frequently targeted pocket.

Another similar approach to FTMAP is pMD (probe-based solvent molecular dynamics simulation) which can detect druggable binding sites in soluble proteins using probe molecules as part of the solvent environment of an MD simulation. This includes cryptic pockets manifested through an induced-fit mechanism. pMD can be considered as the computational equivalent of multi solvent crystallography [22], and fragment-based NMR spectroscopy [23, 24]. This technique considers protein motion directly during the druggable site identification process. It identifies druggable pockets and their maximum affinity [6] by measuring binding free energy between the protein and probe molecules, thus reducing false positives [6, 25, 26]. Furthermore, pMD is distinct from other methods as it does not use surface descriptors and does not rely on a training set. Prakash et al. [6] used pMD to identify interaction hotspots on the surface of oncogenic K-Ras G12D. They identified five druggable sites and three sub-sites. Two sub-sites are classified as part of p1 and p4 pockets, while three druggable sites overlap with p1, p2, and p3 pockets. Furthermore, pMD yielded additional reactive surfaces that could be protein-protein and protein-membrane interaction surfaces. Prakash et al. [27] expanded the scope of pMD to include membrane-bound proteins by modifying the Lennard-Jones (LJ) pairwise nonbonded interactions between selected atoms

of the probe and lipid preventing probe partitioning into a bilayer. Two different modes of membrane-bound orientations of oncogenic G12D and G13D K-Ras were studied and yielded the binding sites that are solvent accessible and not the one occluded by membrane[27]. Sayyed-Ahmad et al.[28] extended the pMD-membrane method by incorporating seven chemically diverse probe types. They also established a method to quantify the density of probes on protein surfaces and construct surface topography maps based on the probe-binding affinity of surface residues to exclude high probe-density spots that do not have pocket-like geometrical characteristics. They also studied oncogenic G12D, G12V and G13D K-Ras mutants in two different membrane orientations. The results robustly identified previously characterized allosteric druggable sites and other reactive surface regions. Finally, Capra et al. [29] developed ConCavity and compared it to both conservation-based and structure-based methods using precision-recall (PR) curves and the Jaccard coefficients. It was found that the ability of structure-based methods to identify ligand binding sites decreases as the number of chains in the structure increases. When there are more than five chains in a protein, conservation alone is more effective than structure-based approaches, and the ConCavity method, which takes advantage of this complementary information, performs very well when there are more than five chains in the protein. Despite ConCavity's excellent performance, some kinds of proteins, such as the ActR protein, did not perform as well. These poor results are due in large part to misleading information about the evolutionary conservation of sequences and ligands binding completely outside well-defined concave surface pockets [29].

In this thesis, we will utilize TACTICS [9] (Trajectory-based Analysis of Conformations to Identify Cryptic Sites) to investigate the binding sites on the surface of three K-Ras variants (G12D, pT32/G12D and C118SOH/G12D). This analysis technique is designed to predict new binding sites based on a complex

machine learning algorithm that utilizes an extended CryptoSite training dataset to analyze the physicochemical properties and geometry of input structures extracted from molecular dynamics trajectories. This approach is more comprehensive than other methods such as ConCavity [29], LIGSITE [30], COACH [31], FTMAP [15], and FPocket [32].This is because they use a single structure as input data limiting their ability to find cryptic sites that are not present in all conformations. It is also more robust than pMD as it utilizes CryptoSite training set as has been demonstrated to be robust in finding druggable pockets in three different proteins: SARS-CoV-2 main protease and methyltransferase and the Yersinia pestis aryl carrier protein.

Chapter 2: Methods

2.1 Molecular Dynamics (MD) simulation

Molecular dynamics is a widely spread computer simulation technique used to analyze protein structure and function. It is also indispensable tool in structural drug design. It is based on applying Newtonian dynamics by measuring the force and acceleration experienced by every atom in a molecular assembly to study their time evolution and hence calculate their equilibrium physicochemical properties. The time evolution of each atom trajectory is obtained by solving

$$m_i \frac{d^2 \vec{r}_i}{dt^2} = -\nabla_i V(\vec{r}_1, \vec{r}_2, ..., \vec{r}_N) \tag{2.1}$$

Where m_i is the mass of each atom (i) at position (r_i) in the system, and V is The potential energy. Some of the commonly used packages for molecular dynamics (MD) simulation of biomolecules are GROMACS, AMBER, and NAMD [33, 34].

In this thesis, we studied three K-Ras variant systems (G12D, pT32/G12D and C118SOH/G12D). The MD simulations of the studied systems have been carried out in previous studies[35, 36] using a standard protocol as follows: All three systems were started from the high-resolution crystal structure of G12D K-Ras (PDB ID: 4DSO). The phosphorylated G12D K-Ras variant (pT32/G12D) was con-

structed by mutating Tyr32 to pTyr32, while the oxidized G12D K-Ras variant (C118SOH/G12D) was constructed using CHARMM (CGenFF) [37, 38] by converting Cys-SH to Cys-SOH. The force field parameters for cysteine sulfenic acid (SOH) were obtained from [39]. The bound guanosine-diphosphate-monothiophosphate (GSP) was removed and replaced with guanosine triphosphate (GTP). All other co-crystals were also removed except for Mg^{2+} ion. By using the PROPKA program, the protonation state of each amino acid residue was predicted[40]. For C118SOH it could be deprotonated (Cys-SO -) because it is a weak acid. C118SOH pKa values range from 5.9 to 7.2 in selected proteins and dipeptides, suggesting that a significant fraction of the molecule exists in the protonated state[39]. C-terminal and anionic residues were deprotonated, while N-terminal and cationic residues were protonated. The resulting structures were then dissolved in a TIP3P cubic box with a buffer spacing of 10 Å to ensure the periodic images of the protein do not interact with the edges of the box. Additionally, (Na^+) and (Cl^-) ions were added to the solvated systems to neutralize and maintain physiological ion concentrations (0.15 M).

For MD simulation runs, an initial conjugate gradient approach was applied with 5000 steps to minimize the system. Using a harmonic force constant ($k = 4.0 kcal/mol Å^2$), the system was heated from $(0 - 310)K$, constrained proteins, and GTP heavy atoms, and removed slowly at a fixed pressure. NPT ensemble with periodic boundary conditions and a 2.0 fs run time along with the SHAKE algorithm for constrained covalent bonds including hydrogen atoms.was also used. In addition, PME methods [41], have been used with a grid density of about 1/Å, between 10 Å and 12 Å, the unbound interactions gradually and smoothly close, and the cutoff occurs at 14Å. NPT ($T = 310K, P = 1.0atm$) simulation, which is consistent with experiments. Langevin dynamics [42] was utilized to regulate temperature and pressure with a damping coefficient of 10 ps^{-1}. Finally, to maintain

constant pressure, the Nose Hoover Langevin piston method [43] was used with a piston period of 200 fs and a decay time interval of 100 fs.

2.2 Trajectory-based Analysis of Conformations to Identify Cryptic Sites (TACTICS)

In this thesis, we aim to identify potential binding sites on the surface of three K-Ras variants and determine the differences between them. We utilized TACTICS[9], a recently developed trajectory-based analysis of MD conformations, to achieve this goal. TACTICS starts by using K-means clustering to select frames from the overall conformational heterogeneity of MD simulation data. It then identifies possible druggable sites in each conformation and integrated these conformational data into a random forest model based on each conformation protein motion and geometry. Finally, TACTICS uses fragment docking to score residues in identified binding pockets. In the following subsections we shed more light on the details of the various stages of TACTICS approach.

2.2.1 Generating a machine learning training database

TACTICS uses a database of 11,201 crystal structures (23% of the proteome) to train a machine learning (ML) model. Each protein in the database has both a holo-structure (with a ligand at the cryptic site) and an apo-structure (without a ligand at the cryptic site). For each holo-structure in the set, all apo-structures with at least 95% sequence identity available in PDB were added. In an attempt to mimic MD trajectories in which many frames may have pockets that are different from cryptic pockets, cryptic pockets are considered to be less hydrophobic and more flexible than other less relevant pockets [44]. PyMOL[45] was subsequently used to select

residues within 9.0 Å of the co-crystallized ligand to align the holo-structure with each extended-database structure. The database was then purged of extended-site structures containing ligands other than water or metals within 5.0 Å of the holo ligand. There is also a variance in the number of apo structures per holo structure. To reduce this issue, a maximum of 50 apo structures per holo-structure is used. This does not completely eliminate the issue that proteins with fewer than 50 structures will be undersampled.

It is worth mentioning that several apo-structures have partially or fully open cryptic sites, making them bindable without ligands. ML models are affected by this shortcoming. In order to reduce the impact of this shortcoming, proteins exist in a spectrum of conformations with cryptic sites that may or may not be accessible. Because of this, deciding whether intermediate states are bindable or unbindable requires arbitrary decisions.

2.2.2 Construction of ML models

An ML model was constructed to identify cryptic sites that are accessible in the input structure. It achieves this task by initially using Ligsite [30] to find out which points near the protein are part of geometric pockets. After that, ConCavity [29] is used to cluster these points into binding sites and assign scores to each residue according to its location. In the ML model, the ConCavity score is calculated from the aligned MD trajectories and reference structures. By measuring the average ConCavity score of the residues to either side of the residue (in sequence), an ML model can detect the number of residues in each druggable cryptic pocket containing more than one residue.

20% of the data was used as testing data, which was used for evaluating the model performance, but not for training the model. The number of true and false

positives varies because of the confidence threshold that the model applies to classify an item as True. To quantify the performance of the ML model, by calculating the receiver operating characteristic (ROC), which is a graph of true positive rate versus false positive rate, as well as AUROC (area under receiver operating characteristic), applied in TACTICS. The rest 80% of the data was used to train a random forest model by using Scikit-learn [46].

2.3 Integrating the ML model into a larger algorithm

To reduce the number of false positives of the ML model, the ML model is incorporated into a multi-step procedure:

- TACTICS algorithm begins with using MD analysis [47, 48] to apply k-means clustering on the aligned MD trajectory.
- ML is then applied to the centroid of each cluster. Thus, the ML model predicts the probability of a residue's presence in a cryptic site.
- A set of predicted binding regions are assigned according to predicted probabilities that are compared with a user-specified threshold to reduce the number of false positives. The algorithm yields no predictions if the standard deviation of ML scores among all clusters of the trajectory is less than a certain threshold.
- Finally, fragment docking (AutoDock Vina) is used to investigate which predicted pockets are potentially druggable. The "dock score" for each residue in Cryptosite is a calculation based on how many docked fragment ligands are within 3.5 Å distance as shown in figure 2.1.

The TACTICS output is displayed using pyMOL[45], where the ML high score residues appear as black or wight sticks, and fragment docking scores appear as

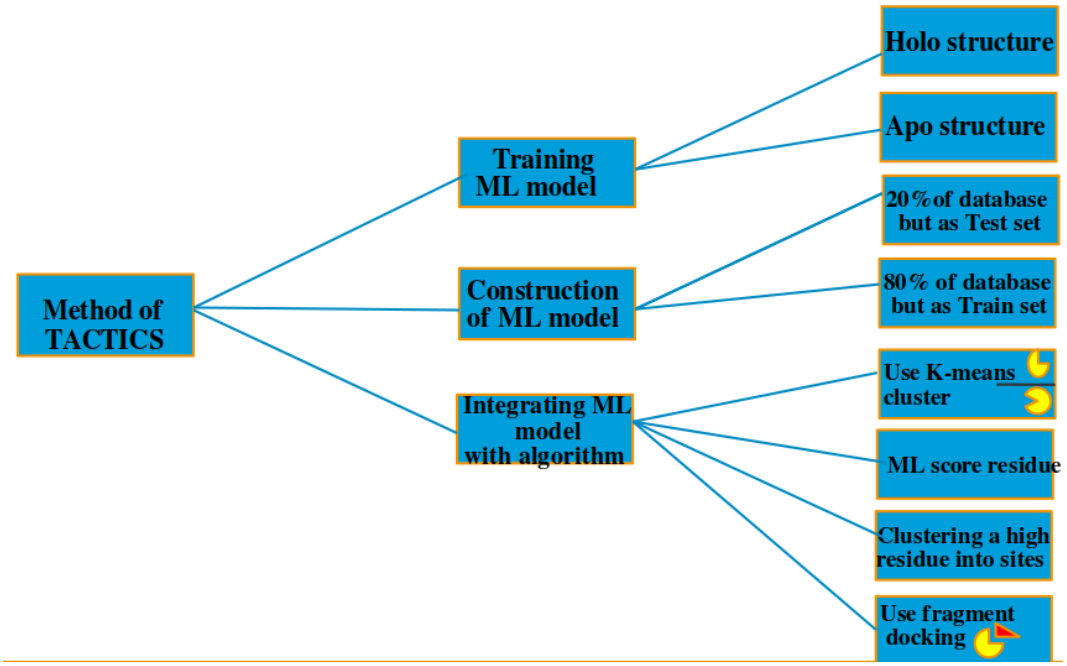

Figure 2.1: TACTICS methods scheme (describing the TACTICS algorithm)

B-factor putty. High confidence TACTICS predictions are obtained when residues with high ML scores are in the same region as those with high docking scores as shown in Figure 2.2.

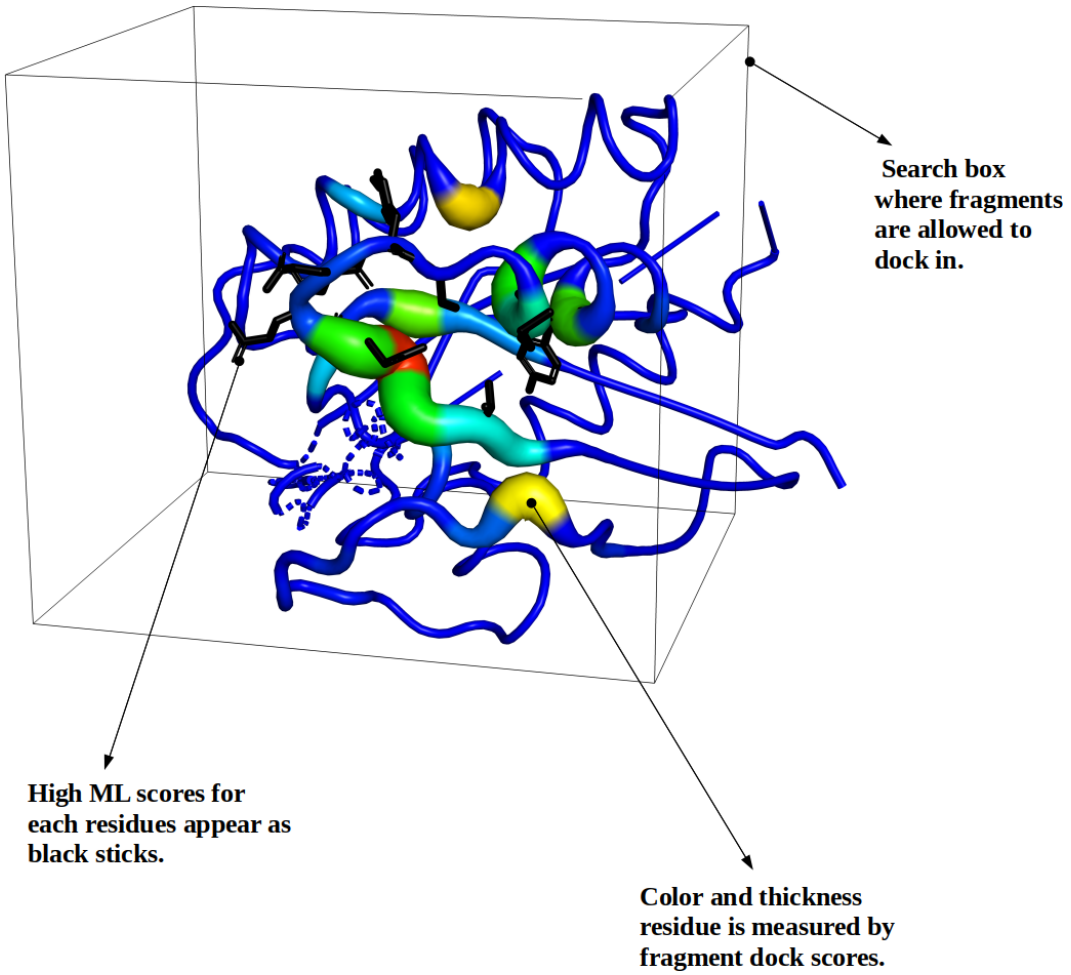

Figure 2.2: A visualization of TACTICS output in PyMOL. ML predictions are shown as black sticks. Fragment docking predictions are stored as B-factor values and shown in putty representation.

Chapter 3: Results

In this work, we applied TACTICS on three different K-Ras variants, G12D, C118SOH/G12D, and pT32/G12D, to identify potential druggable binding sites on the protein surface. In the following sections, first we will show how we calibrated different TACTICS parameters, then we will present our results and predictions for each K-Ras variant.

3.1 Calibration of TACTICS Parameters

To begin with, for TACTICS to work properly, protein trajectory should be aligned using CA atoms. All residue names must also be one of the standard amino acids, and every residue must have an alpha carbon(CA). Atoms that do not meet these requirements are deleted. Additionally, the first amino acids should have chain A[9]. Since K-Ras has a bound nucleotide and we are not interested in including the nucleotide-binding pocket as a prediction, all nucleotide atom names are replaced by CA.

There are three adjustable parameters in TACTICS that need to be carefully assigned. These include the number of clusters ($N_{clusters}$), machine learning score minimum (ϵ_{ml}), and threshold and maximum frame-to-frame standard deviation (σ_{ml}). $N_{clusters}$ determines the number of conformations sampled from the frames of the aligned MD trajectory. It usually ranges from 3 to 15. $N_{clusters}$ also limits the

number of TACTICS predictions. However, if pocket-less frames are present, the number of TACTICS predictions will be less than $N_{clusters}$. ϵ_{ml} is the minimum ML score required to cluster residues into binding sites for each frame. It ranges from 0.5 to 0.9. σ_{ml} is the maximum frame-to-frame standard deviation residue scores must have when clustering residues into binding sites. This means that residues that are assigned in few frames, will be discarded [9]. Choosing these parameters is important to detect binding pockets accurately. Using large values will yield no results while choosing small values will yield spurious results. In this study, we used $N_{clusters}$ = 10, ϵ_{ml} = 0.6, σ_{ml} = 0.1875. T

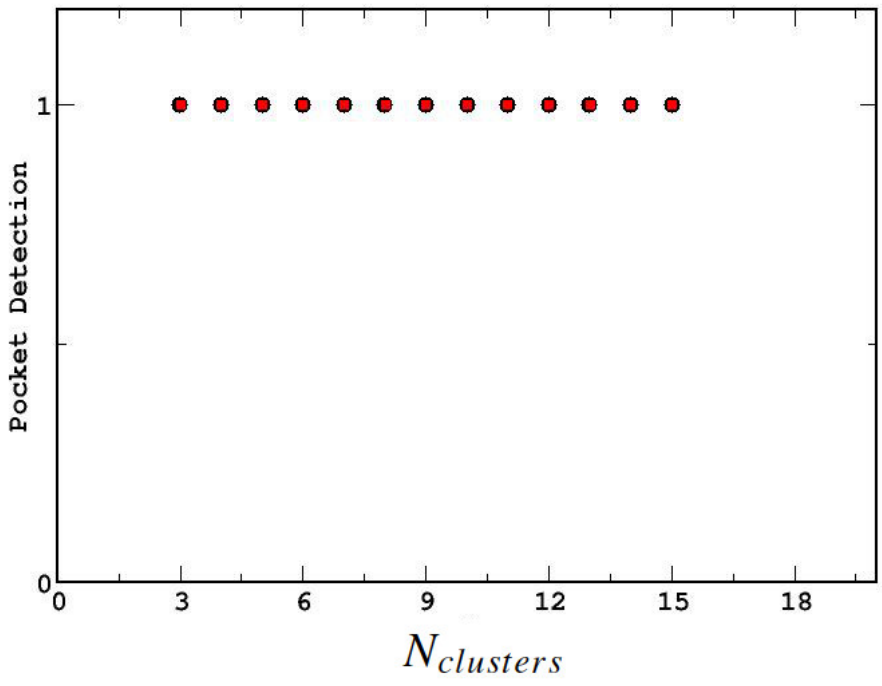

Figure 3.1: Pocket detection as a function of $N_{clusters}$. Pocket p4 appears in the G12D variant (black circles), while pocket p2 appears in C118SOH/G12D variant (red squares).

To make sure that our choice of TACTICS parameters is optimal, a systematic sensitivity study of the three parameters was carried out to determine their effect on the detection of K-Ras binding pockets in the two variants G12D and

C118SOH/G12D. Figure 3.1 suggests that the number of clusters does not affect the detection of pocket p2 in C118SOH/G12D and pocket p4 in G12D. This is reasonable as the number of known binding sites on the surface of K-Ras is not large. Based on this analysis, we chose an optimal $N_{clusters}$ value of 10.

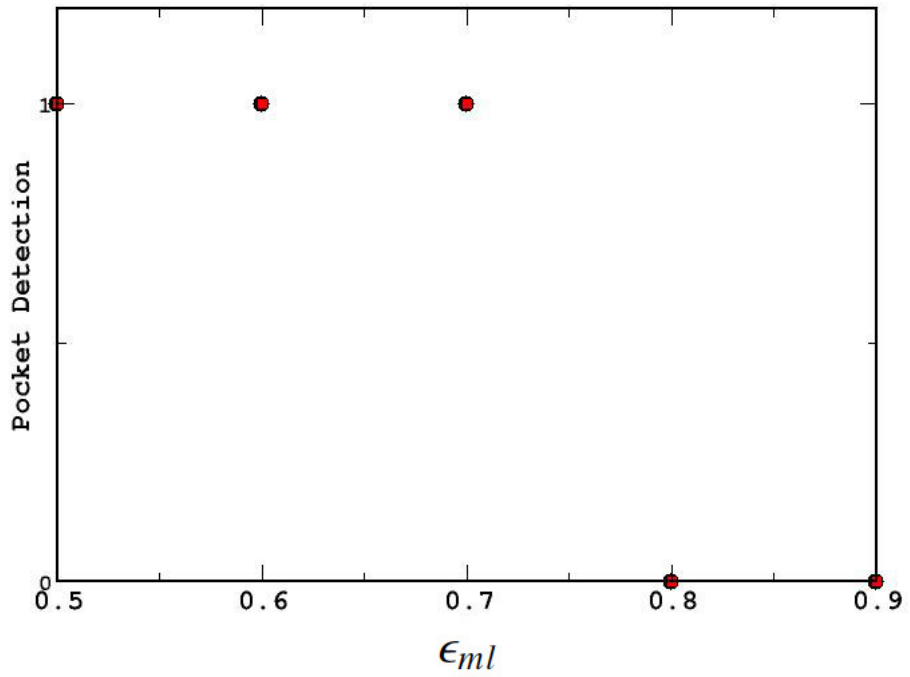

Figure 3.2: Pocket detection as a function ϵ_{ml}. Pocket p4 appears in the G12D variant (black circles), while pocket p2 appears in C118SOH/G12D variant (red squares).

Figure 3.2 shows that ϵ_{ml} significantly affect the detection of pocket p2 in C118SOH/G12D and pocket p4 in G12D. It is clear that ϵ_{ml} values below 0.7 can detect both pocket p4 in C118SOH/G12D and p2 in G12D. To avoid having false positive detection, we chose an optimal ϵ_{ml} of 0.6.

Finally, Figure 3.3 shows that σ_{ml} also affect the detection of pocket p2 in C118SOH/G12D and pocket p4 in G12D. It is clear that σ_{ml} values below approximately 0.2 can detect pocket p4 in G12D, while values below approximately 0.32 can detect p2 in C118SOH/G12D. To detect both pockets and avoid having

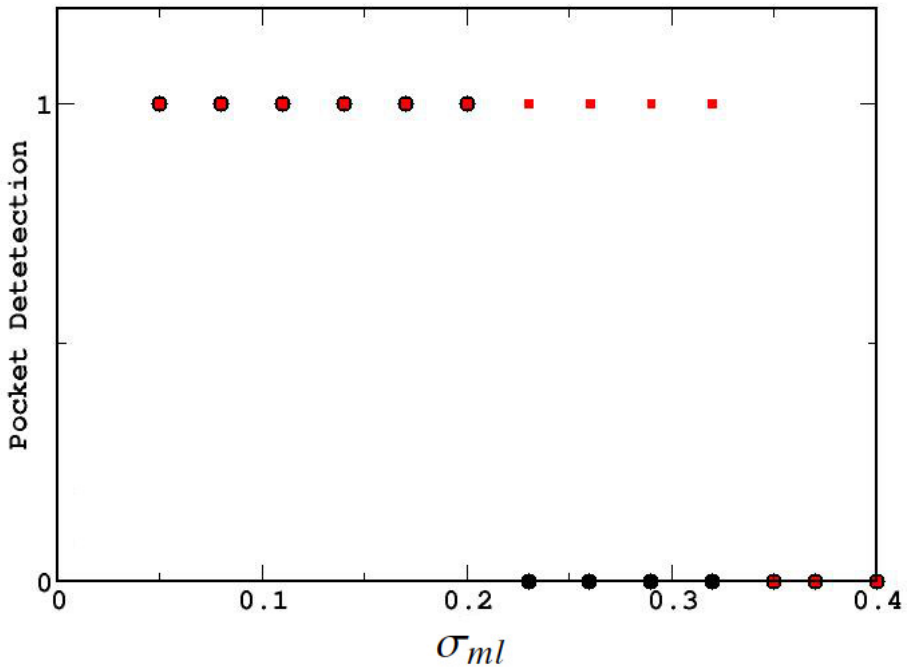

Figure 3.3: Pocket detection as a function σ_{ml}. Pocket p4 appears in the G12D variant (black circles), while pocket p2 appears in C118SOH/G12D variant (red squares).

false positive detection, we chose an optimal σ_{ml} of 0.1875.

3.2 Analysis of Binding Sites in G12D K-Ras variant

Mutations in the Ras protein occur in three regions: the P loop (phosphate-binding loop, residues 10-17) and the switch regions: SI (residues 30-38) and SII (residues 60-76). In addition, SI and SII undergo large conformational changes upon GDP/GTP exchange or state 1/state 2. Conformational changes between the two states indicate a common mechanistic basis inherent in the high flexibility of the switch regions. Shima et al[49] investigated different Ras mutations such as H-RasT35S-GppNHP to clarify the mechanism of state transitions. H-RasT35S has

Cluster	p1	p2	p3	p4
1	NA	NA	109, 111	NA
2	NA	NA	NA	29, 31 35
3	NA	NA	NA	21, 31, 35
4	71	NA	NA	29, 31, 35, 57
5	NA	NA	NA	29, 31, 35
6	NA	NA	106, 109, 110	17, 35, 57
7	NA	NA	NA	17,31
8	72	60	NA	NA

Table 3.1: TACTICS predictions for residues as part of a binding site for each conformational cluster of G12D variant.

a reduced affinity for Ras effector proteins without Thr35 being involved in any interactions [49, 50].

The ^{31}P NMR spectra of H-RasT35S-GppNHP showed that SI is shifted away from the guanine nucleotide and conformational instability of the SI loop and this is because of the loss of the direct and Mg^{+2}-coordinated indirect interactions between Thr35 of H-Ras and the γ-phosphate of GppNHp, and revealed two surface pockets (pocket p1 and pocket p2) in state 1, which are absent in state 2. Shima et al [49] found that the transition from state 2 to state 1 is required for the dissociation of the two loops(SI and SII) of GppNHP. Namely, by dissociating The35 and Gly60 from the phosphate resulting in heavy hydrogen-exhausting bond network transitions. For the studied transition from state 1 to state 2, they used the crystal structure M-RasP40D-GppNHp, which requires the binding of Thr45 and Gly70 toγ-phosphate. Further studies of Ras state transitions will improve our ability to develop drugs targeting Ras proteins and other small GTPases[11, 49, 51].

Our TACTICS analysis showed that optimal parameters yield the presence of pocket p4 (see Figure 3.4, residues: 17, 21, 29, 31, 35) in many conformational clusters of the G12D variant. Pocket p3 (see Figure 3.5), residue: 106, 109-111) were also present. However, pockets p1 and p2 were not detected at all (see table 3.1).

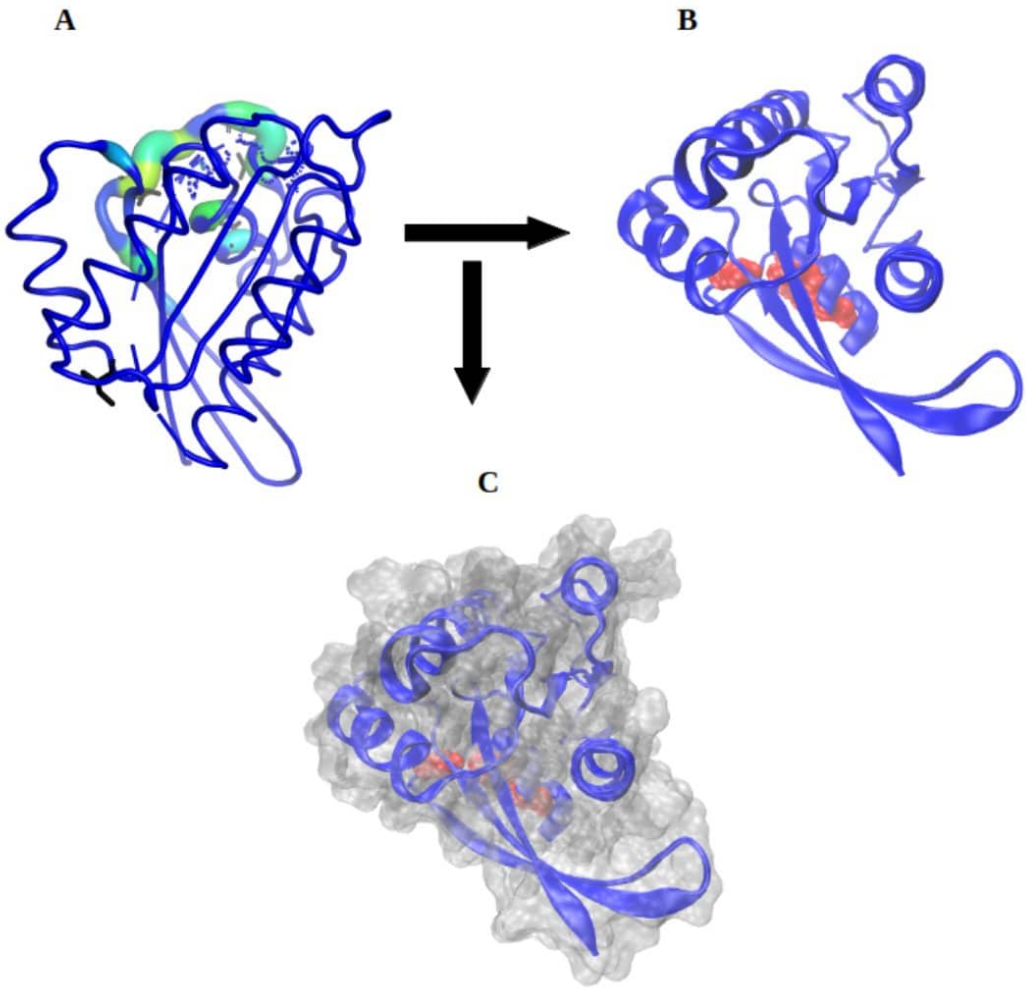

Figure 3.4: A) TACTICS predictions of G12D binding site residues are shown as B factor putty obtained from fragment docking and black sticks obtained from machine learning. B) Pocket p4 is shown as a red surface, while the protein is displayed as a cartoon. C) The same as with transparent protein surface overlay.

This suggests that for the non-oxidized G12D variant pocket p4 can be selectively targeted.

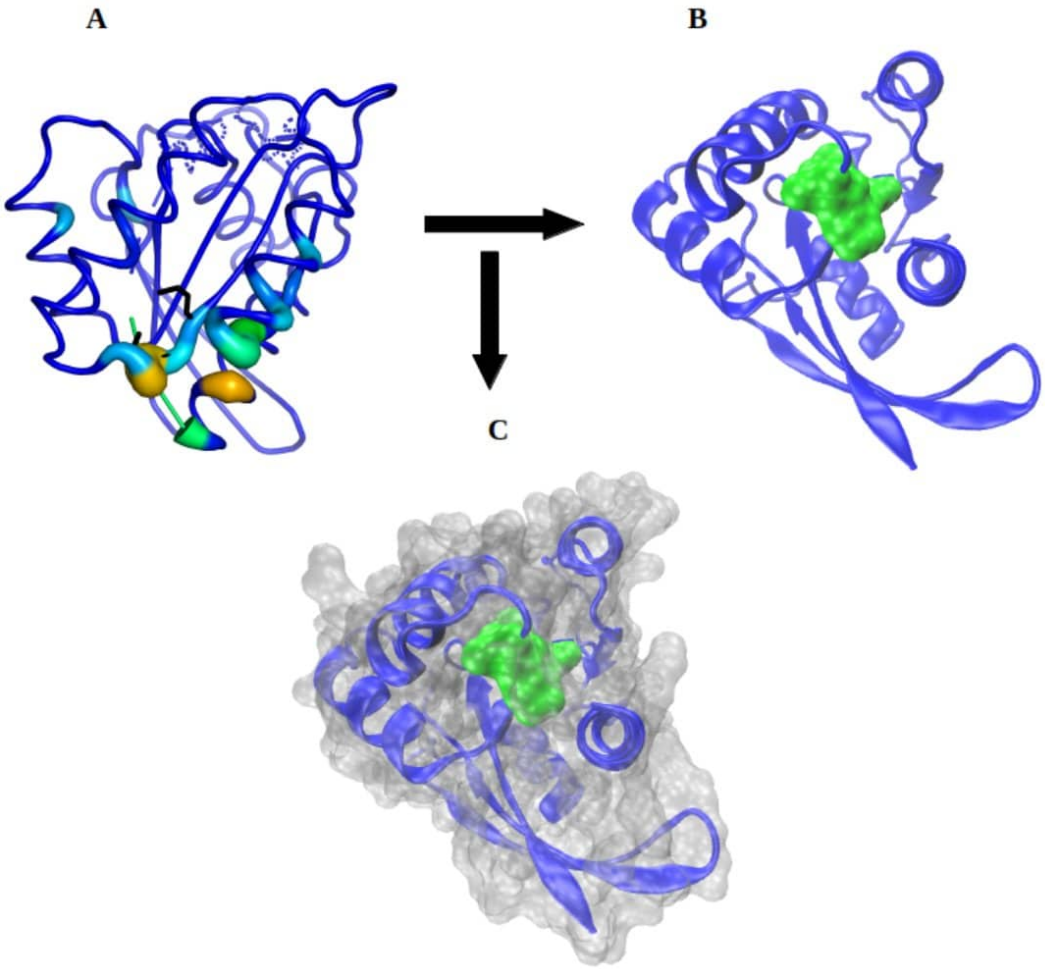

Figure 3.5: A) TACTICS predictions of G12D binding site residues are shown as B factor putty obtained from fragment docking and black sticks obtained from machine learning. B) Pocket p3 is shown as a green surface, while the protein is displayed as a cartoon. C) The same as with transparent protein surface overlay.

3.3 Analysis of Binding Sites in oxidized C118SOH/G12D K-Ras variant

There are many studies investigating the implications of oxidation of Cysteine118 in the three isoforms of Ras protein (N-Ras, K-Ras,H-Ras) as well as the change in dynamics, structural, and conformational properties of the nucleotide-binding sites. A recent MD study [36] suggested that switch SI and SII are more dynamic in the oxidized C118SOH/G12D variant.This increase in flexibility could facilitate GEF

Cluster	p1	p2	p3	p4
1	NA	NA	110, 109, 139	NA
2	NA	60-65,99	NA	NA
3	67	61	NA	NA
4	67,71	61,62	NA	NA
5	67	60-66, 92	NA	NA
6	NA	NA	109,110, 137	NA
7	67, 71	61-65	NA	NA
8	N/A	61, 62	NA	NA
9	NA	NA	NA	NA

Table 3.2: TACTICS predictions for residues as part of a binding site for each conformational cluster of C118SOH/G12D variant.

binding and/or transmission between active and inactive states.

Our TACTICS analysis with optimal parameters showed the presence of pocket p2 in many conformational clusters of the C118SOH/G12D variants (see Figure 3.6, residues: 61-65,92,99). It also showed the presence of pocket p3 (see Figure 3.7, residues: 107-110, 137-139), but to a lesser extent than p2. Only two residues (residue: 67,71) that belong to pocket p1 appeared and no residues of pocket p4 were detected (see table3.2). These results suggest that in the oxidized C118SOH/G12D variant switch SII region is more dynamic than that of G12D. These results suggest pocket p2 might be the more suitable target to inhibit the active oxidized variant. Some known putative drugs that bind to these pockets are reported in the literature [19].

3.4 Analysis of Binding Sites in phosphorylated pT32/G12D K-Ras variant

Many studies have indicated that Tyr32 influences the GTPase activity and effector binding by altering the conformation of Ras protein, and these studies have demonstrated that the change in Tyr32 orientation is associated with the effect on canonical

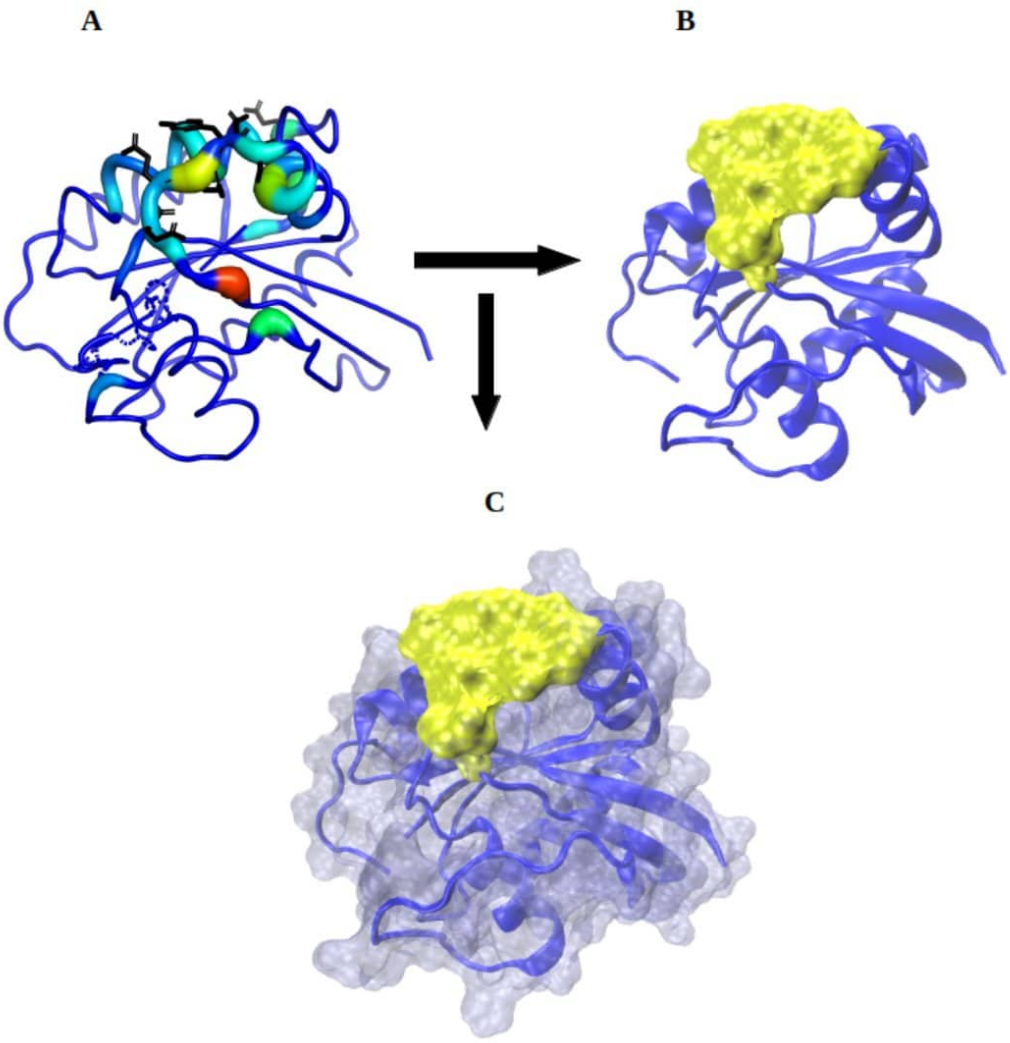

Figure 3.6: A) TACTICS predictions of C118SOH/G12D binding site residues are shown as B factor putty obtained from fragment docking and black sticks obtained from machine learning. B) Pocket p2 is shown as a yellow surface, while the protein is displayed as a cartoon. C) The same as with transparent protein surface overlay.

switches, as well as the consequent possibility of determining the active and inactive conformation of Ras protein[52–54].

The Tyr32 residue can be phosphorylation via scr-protein tyrosine phosphate. This phosphorylation state induces conformational changes in switch regions due to the additional electrostatic repulsion between the negatively charged Asp38 and Asp57 residues within the nucleotide-binding site. It most likely traps K-Ras into an inactive GTP-bound state and hence decreases its affinity towards effector

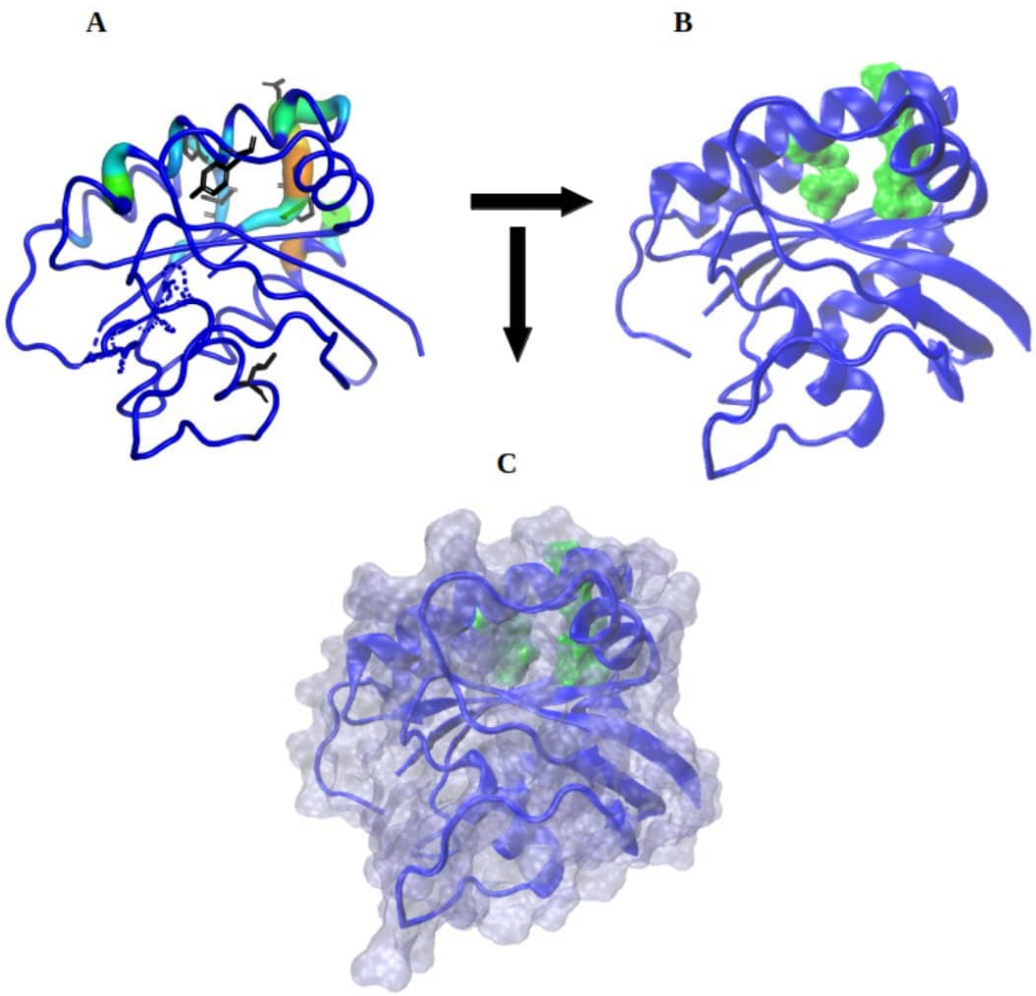

Figure 3.7: A) TACTICS predictions of C118SOH/G12D binding site residues are shown as B factor putty obtained from fragment docking and black sticks obtained from machine learning. B) Pocket p3 is shown as a green surface, while the protein is displayed as a cartoon. C) The same as with transparent protein surface overlay.

protein Raf. Dephosphorylation Tyr32 via SHP2 protein tyrosine phosphate conversely enhances Raf binding. This suggests that the phosphorylation state of Tyr32 influences the GTPase cycle. As a result, disrupting the balance between these processes may have adverse functional effects that lead to cancer [55, 56].

MD simulations were used to examine the structural and dynamical changes caused by the phosphorylation of Tyr32 in G12D-K-Ras [35]. These simulations indicate that the switch SI region is more flexible in pT32/G12D than its

Cluster	p1	p2	p3	p4
1	NA	NA	107-109	NA
2	NA	NA	164	N/A
3	NA	NA	107-109, 137-139, 162, 164	36, 57
4	NA	NA	107-09, 137, 139	N/A
5	NA	NA	101, 107-111, 137-139	N/A
6	NA	NA	107-111, 137, 139	N/A
7	NA	NA	NA	NA
8	54,55	NA	NA	NA
9	NA	NA	109,110, 138, 139	NA

Table 3.3: TACTICS predictions for residues as part of a binding site for each conformational cluster of pT32/G12D variant.

dephosphorylated counterpart, whilst the switch SII has a reduced degree of flexibility. This suggests the phosphorylation process affects the binding of Ras with its effectors.

Similar to the previous systems, we used the same parameters to identify the druggable binding sites in pT32/G12D variant. It also showed the presence of pocket p3 (see Figure 3.8, residues: 107-110, 137-139). Only two residues(see Figure 3.9, residues: 67,71) that belong to pocket p1 appeared and no residues of pocket p4 were detected. Residues 54 and 55 which are part of pocket p1 rarely appeared. There were also rare occurrences of residues 46 and 49) which belong to β2 and β3 strands, respectively.

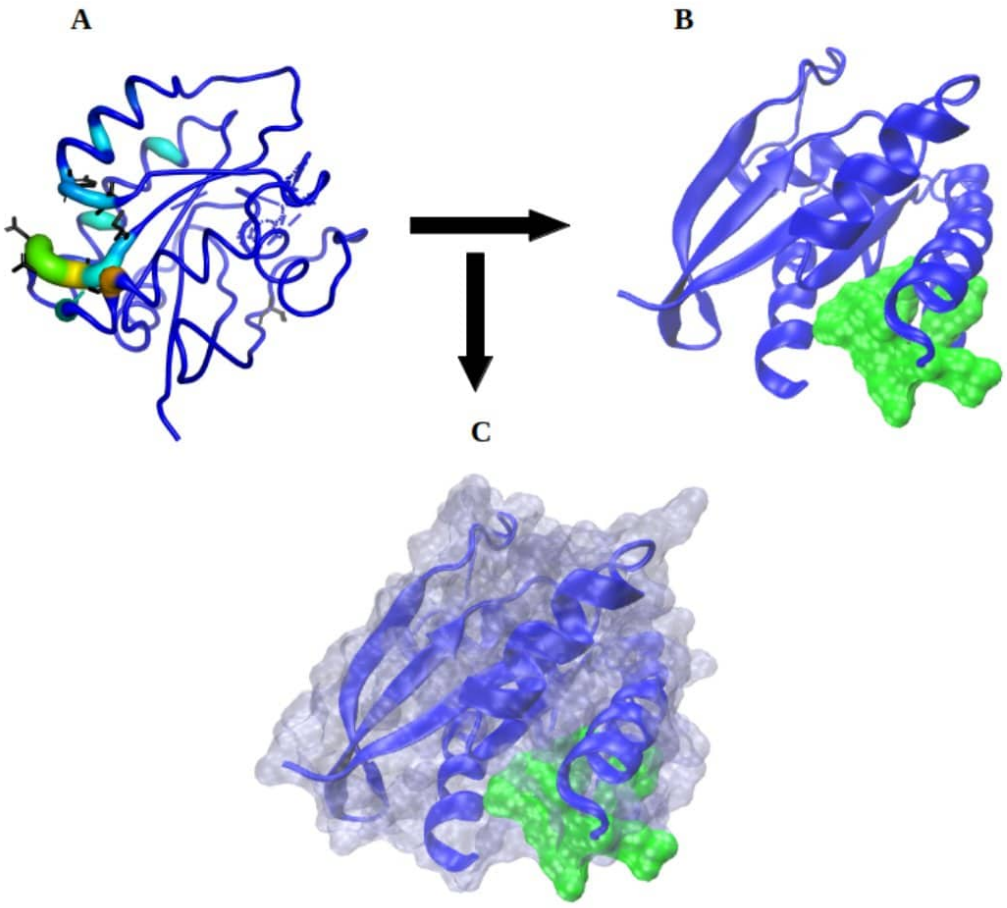

Figure 3.8: A) TACTICS predictions of pT32/G12D binding site residues are shown as B factor putty obtained from fragment docking and black sticks obtained from machine learning. B) Pocket p3 is shown as a green surface, while the protein is displayed as a cartoon. C) The same as with transparent protein surface overlay.

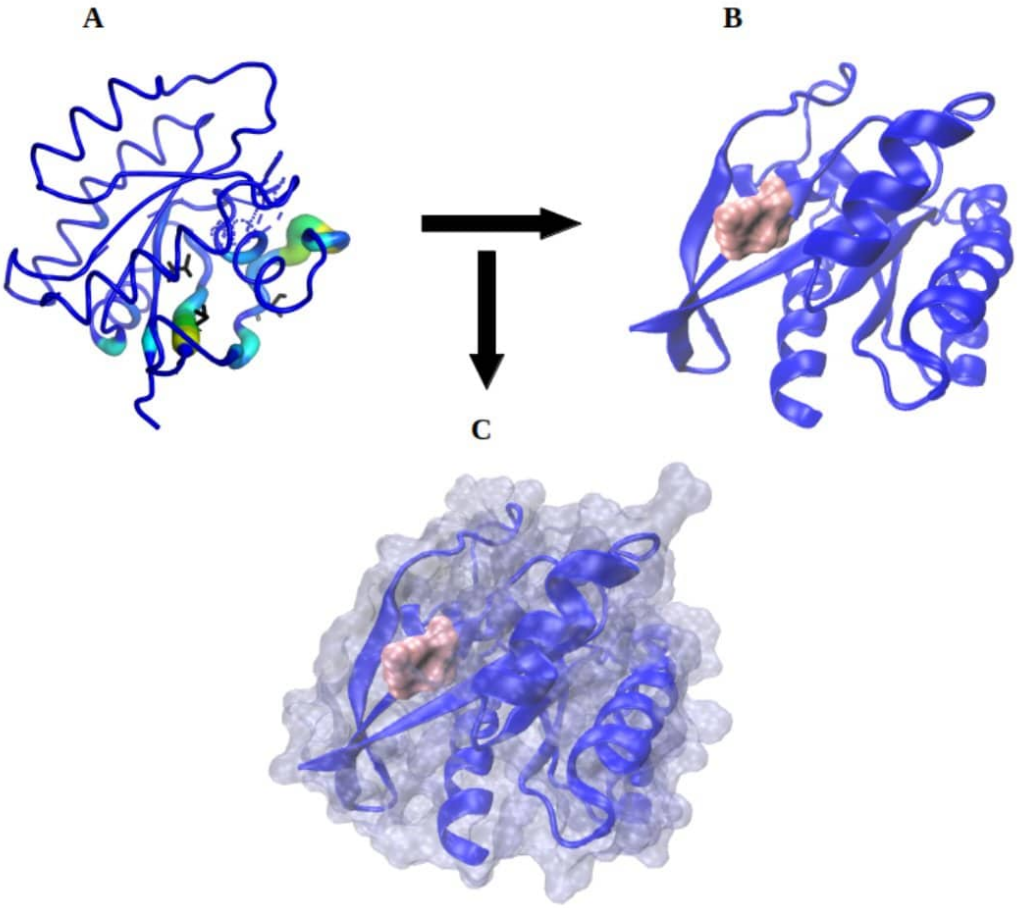

Figure 3.9: A) TACTICS predictions of pT32/G12D binding site residues are shown as B factor putty obtained from fragment docking and black sticks obtained from machine learning. B) Pocket p1 is shown as a pink surface, while the protein is displayed as a cartoon. C) The same as with transparent protein surface overlay.

Chapter 4: Conclusions

The aim of this thesis was to identify drug-binding sites in proteins utilizing MD simulation data. This is done using a novel approach that combines machine learning and a complex algorithm called TACTICS. As a starting point, K-means clustering was applied to select frames, and a random forest model was developed for each conformation to predict potential binding residues. Finally, fragment docking was used to investigate which predicted pockets could be druggable. Accurate predictions are made when a high ML score is in the same region as a residue with a high docking score is obtained [9].

K-Ras protein is one of the small GTPase proteins. It is related to major biological functions in the human body, such as controlling the growth, proliferation, differentiation, and apoptosis of cells. As a result of that, any mutation in the K-Ras protein can lead to cancer. One of the frequent mutations in K-Ras protein is G12D, which causes different types of cancer, including pancreatic cancer. Recently, several approaches have been developed for finding proteins' druggable binding sites with the aim of developing an effective drug. For this aim, we applied TACTICS to three different types of K-Ras variants: G12D, C118SOH/G12D and pT32/G12D.

TACTICS has shown that K-Ras pocket p3 exists in the three systems, with a particular appearance in pT32/G12D, where it occurs in many different MD sampled conformations. The ubiquitous appearance of p3 indicates that it is a major

druggable pocket. Pocket p4 was most prominent in G12D variant, while pocket p2 appeared in the C118SOH/G12D variant system. These results demonstrate the robustness of TACTICS in identifying previously identified K-Ras binding sites and its ability to differentiate between their appearance in different K-Ras variants.

References

(1) Mansoori, B.; Mohammadi, A.; Davudian, S.; Shirjang, S.; Baradaran, B. *Advanced pharmaceutical bulletin* **2017**, *7*, 339.
(2) Gimple, R. C.; Wang, X. *Frontiers in oncology* **2019**, *9*, 965.
(3) McCormick, F. *Journal of molecular medicine* **2016**, *94*, 253–258.
(4) Bos, J. L. *Cancer research* **1989**, *49*, 4682–4689.
(5) McCarthy, M.; Prakash, P.; Gorfe, A. A. *Acta biochimica et biophysica Sinica* **2016**, *48*, 3–10.
(6) Prakash, P.; Hancock, J. F.; Gorfe, A. A. *Proteins: Structure, Function, and Bioinformatics* **2015**, *83*, 898–909.
(7) Sayyed-Ahmad, A.; Prakash, P.; Gorfe, A. A. *Proteins: Structure, Function, and Bioinformatics* **2017**, *85*, 1618–1632.
(8) Uprety, D.; Adjei, A. A. *Cancer treatment reviews* **2020**, *89*, 102070.
(9) Evans, D. J.; Yovanno, R. A.; Rahman, S.; Cao, D. W.; Beckett, M. Q.; Patel, M. H.; Bandak, A. F.; Lau, A. Y. *Journal of chemical information and modeling* **2021**, *61*, 2897–2910.
(10) Kessler, D.; Gmachl, M.; Mantoulidis, A.; Martin, L. J.; Zoephel, A.; Mayer, M.; Gollner, A.; Covini, D.; Fischer, S.; Gerstberger, T., et al. *Proceedings of the National Academy of Sciences* **2019**, *116*, 15823–15829.
(11) Pantsar, T. *Computational and structural biotechnology journal* **2020**, *18*, 189–198.
(12) Lu, S.; Ni, D.; Wang, C.; He, X.; Lin, H.; Wang, Z.; Zhang, J. *ACS Catalysis* **2019**, *9*, 7188–7196.
(13) Muratcioglu, S.; Chavan, T. S.; Freed, B. C.; Jang, H.; Khavrutskii, L.; Freed, R. N.; Dyba, M. A.; Stefanisko, K.; Tarasov, S. G.; Gursoy, A., et al. *Structure* **2015**, *23*, 1325–1335.
(14) Anderson, A. C. *Chemistry & biology* **2003**, *10*, 787–797.
(15) Ngan, C. H.; Bohnuud, T.; Mottarella, S. E.; Beglov, D.; Villar, E. A.; Hall, D. R.; Kozakov, D.; Vajda, S. *Nucleic acids research* **2012**, *40*, W271–W275.
(16) Ivetac, A.; Andrew McCammon, J. *Chemical biology & drug design* **2010**, *76*, 201–217.
(17) Kozakov, D.; Grove, L. E.; Hall, D. R.; Bohnuud, T.; Mottarella, S. E.; Luo, L.; Xia, B.; Beglov, D.; Vajda, S. *Nature protocols* **2015**, *10*, 733–755.
(18) Hall, D. R.; Kozakov, D.; Vajda, S. In *Computational drug discovery and design*; Springer: 2012, pp 13–27.

(19) Grant, B. J.; Lukman, S.; Hocker, H. J.; Sayyah, J.; Brown, J. H.; McCammon, J. A.; Gorfe, A. A. *PloS one* **2011**, *6*, e25711.
(20) Harris, R.; Olson, A. J.; Goodsell, D. S. *Proteins: structure, function, and bioinformatics* **2008**, *70*, 1506–1517.
(21) Hetényi, C.
(22) Mattos, C.; Ringe, D. *Nature biotechnology* **1996**, *14*, 595–599.
(23) Shima, F.; Yoshikawa, Y.; Ye, M.; Araki, M.; Matsumoto, S.; Liao, J.; Hu, L.; Sugimoto, T.; Ijiri, Y.; Takeda, A., et al. *Proceedings of the National Academy of Sciences* **2013**, *110*, 8182–8187.
(24) Hajduk, P. J.; Huth, J. R.; Fesik, S. W. *Journal of medicinal chemistry* **2005**, *48*, 2518–2525.
(25) Seco, J.; Luque, F. J.; Barril, X. *Journal of medicinal chemistry* **2009**, *52*, 2363–2371.
(26) Bakan, A.; Nevins, N.; Lakdawala, A. S.; Bahar, I. *Journal of chemical theory and computation* **2012**, *8*, 2435–2447.
(27) Prakash, P.; Sayyed-Ahmad, A.; Gorfe, A. A. *PLoS computational biology* **2015**, *11*, e1004469.
(28) Sayyed-Ahmad, A.; Gorfe, A. A. *Journal of chemical theory and computation* **2017**, *13*, 1851–1861.
(29) Capra, J. A.; Laskowski, R. A.; Thornton, J. M.; Singh, M.; Funkhouser, T. A. *PLoS computational biology* **2009**, *5*, e1000585.
(30) Hendlich, M.; Rippmann, F.; Barnickel, G. *Journal of Molecular Graphics and Modelling* **1997**, *15*, 359–363.
(31) Yang, J.; Roy, A.; Zhang, Y. *Bioinformatics* **2013**, *29*, 2588–2595.
(32) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. *BMC bioinformatics* **2009**, *10*, 1–11.
(33) Patodia, S.; Bagaria, A.; Chopra, D. *Journal of Physical Chemistry & Biophysics* **2014**, *4*, 1.
(34) Karplus, M.; McCammon, J. A. *Nature structural biology* **2002**, *9*, 646–652.
(35) Khaled, M.; Gorfe, A.; Sayyed-Ahmad, A. *The Journal of Physical Chemistry B* **2019**, *123*, 7667–7675.
(36) Shimaa, S. Invwestigating the Effects of Cysteine-118 Oxidation on G12D K-RAS Structure: Insights from MD Simulations, unpublished, 2021.
(37) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. *Journal of computational chemistry* **2008**, *29*, 1859–1865.
(38) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I., et al. *Journal of computational chemistry* **2010**, *31*, 671–690.
(39) Heppner, D. E.; Dustin, C. M.; Liao, C.; Hristova, M.; Veith, C.; Little, A. C.; Ahlers, B. A.; White, S. L.; Deng, B.; Lam, Y.-W., et al. *Nature communications* **2018**, *9*, 1–11.
(40) Olsson, M. H.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. *Journal of chemical theory and computation* **2011**, *7*, 525–537.

(41) Darden, T.; York, D.; Pedersen, L. *The Journal of chemical physics* **1993**, *98*, 10089–10092.
(42) Pastor, R. In *The Molecular Dynamics of Liquid Crystals*; Springer: 1994, pp 85–138.
(43) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. *The Journal of chemical physics* **1995**, *103*, 4613–4621.
(44) Cimermancic, P.; Weinkam, P.; Rettenmaier, T. J.; Bichmann, L.; Keedy, D. A.; Woldeyes, R. A.; Schneidman-Duhovny, D.; Demerdash, O. N.; Mitchell, J. C.; Wells, J. A., et al. *Journal of molecular biology* **2016**, *428*, 709–719.
(45) Schrodinger, L. *Version* **2010**, *1*, 0.
(46) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V., et al. *the Journal of machine Learning research* **2011**, *12*, 2825–2830.
(47) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. *Journal of computational chemistry* **2011**, *32*, 2319–2327.
(48) Gowers, R. J.; Linke, M.; Barnoud, J.; Reddy, T. J. E.; Melo, M. N.; Seyler, S. L.; Domanski, J.; Dotson, D. L.; Buchoux, S.; Kenney, I. M., et al. *MDAnalysis: a Python package for the rapid analysis of molecular dynamics simulations*; tech. rep.; Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2019.
(49) Shima, F.; Ijiri, Y.; Muraoka, S.; Liao, J.; Ye, M.; Araki, M.; Matsumoto, K.; Yamamoto, N.; Sugimoto, T.; Yoshikawa, Y., et al. *Journal of Biological Chemistry* **2010**, *285*, 22696–22705.
(50) Spoerner, M.; Herrmann, C.; Vetter, I. R.; Kalbitzer, H. R.; Wittinghofer, A. *Proceedings of the National Academy of Sciences* **2001**, *98*, 4944–4949.
(51) Ye, M.; Shima, F.; Muraoka, S.; Liao, J.; Okamoto, H.; Yamamoto, M.; Tamura, A.; Yagi, N.; Ueki, T.; Kataoka, T. *Journal of Biological Chemistry* **2005**, *280*, 31267–31275.
(52) Milburn, M. V.; Tong, L.; DeVos, A. M.; Brünger, A.; Yamaizumi, Z.; Nishimura, S.; Kim, S.-H. *Science* **1990**, *247*, 939–945.
(53) Geyer, M.; Schweins, T.; Herrmann, C.; Prisner, T.; Wittinghofer, A.; Kalbitzer, H. R. *Biochemistry* **1996**, *35*, 10308–10320.
(54) Baussand, J.; Kleinjung, J. *Journal of chemical theory and computation* **2013**, *9*, 738–749.
(55) Bunda, S.; Burrell, K.; Heir, P.; Zeng, L.; Alamsahebpour, A.; Kano, Y.; Raught, B.; Zhang, Z.-Y.; Zadeh, G.; Ohh, M. *Nature communications* **2015**, *6*, 1–12.
(56) Kano, Y.; Cook, J. D.; Lee, J. E.; Ohh, M. In *Seminars in cell & developmental biology*, 2016; Vol. 58, pp 70–78.

Appendix A:

For TACTICS to work properly, the protein trajectory should be aligned using CA atoms. All residue names must also be one of the standard amino acids, and every residue must have an alpha carbon (CA). Atoms that do not meet these requirements are deleted. Additionally, the first amino acids should have chain A. To do that open the PDB file for the G12D K-Ras variant on the terminal using this command:

- $ nano KG12D.pdb
- $ sed 's/MET B/MET A/' KG12D.pdb

Then change the chain name to chain A manually, and do the same for the phosphorylated pT32/G12D K-Ras variant and oxidized C118SOH/G12DK-Ras variant.

Sometimes we need to change the orientation of the protein structure to show the pockets by using the VMD program:

- - set sel [atomselect top "all"]
- - set com [measure center $sel weight mass]
- - set matrix [transaxis x -180]
- $ sel moveby [vecscale -1.0 $ com]
- $ sel move $ matrix
- $ sel moveby $ com