Politecnico di Torino

Master's degree in Biomedical Engineering

Molecular Dynamics of the Alsin DH/PH domain toward a better understanding of Infantile-onset Ascending Hereditary Spastic Paraplegia

Supervisor

Prof. Marco Agostino Deriu

Candidate

Marco Cannariato

Co-supervisor

Marcello Miceli

Academic year 2020/2021

Contents

Abstract

Infantile-onset ascending hereditary spastic paralysis (IAHSP) is rare neurodegenerative disease characterized by onset of spasticity to lower limbs within the second year of life and progression towards spastic tetraparesis. This disorder is associated with mutations at the Amyotrophic Lateral Sclerosis type 2 (ALS2) gene, which encodes for Alsin, a protein composed by 1657 amino acids organized in multiple domains. Several studies on transgenic mice have highlighted its crucial role in vesicular trafficking, neuronal development, and homeostasis by virtue of its ability to interact with two guanosine triphosphatases, Rac1 and Rab5. In particular, evidence suggest that Rac1 can bind Alsin central region, composed by two structured domains i.e. a Dbl Homology (DH) domain followed by a Pleckstrin Homology (PH) domain. *In vitro* experiments have shown that this interaction is necessary for the subsequent activation of Rab5 through Alsin C-terminal region, leading to the maturation and fusion of different types of vesicles. However, as far as we know, the three-dimensional structure of Alsin protein and its relationship with specific functions are still unknown. Computational Molecular Modelling is an elective tool to study nanoscale level biological systems, both allowing to model the 3D structure and the dynamics of proteins. In this work, the first homology model of Alsin DH/PH domain was developed and studied through Molecular Dynamics both in presence and in absence of its binding partner, Rac1. As a proof of the results robustness, the employed experimental setup was first validated replicating MD studies on homologues DH/PH domains reported in literature. Regarding Alsin DH/PH essential dynamics, it consisted in a collective motion of PH region independently of Rac1 interaction. Due to different conformations of DH domain, the presence of Rac1 seems to stabilize an open state of the protein, while absence of its binding partner results in closed conformations. Furthermore, Rac1 interaction was able to reduce the fluctuations in the second conserved region of DH motif, which may be involved in the formation of a homodimer. Moreover, the dynamics of DH/PH was described through a Markov State Model to study the pathways linking the open and closed states. In conclusion, this work provided the first all atom model for DH/PH domain of Alsin protein, moreover, MD investigations suggested underlying molecular mechanisms in the signal transduction between Rac1 and Alsin, providing the basis for a deeper understanding of the whole structure-function relationship for the Alsin protein.

1 Introduction

Hereditary Spastic Paraplegia (HSP) is a group of hereditary neurological disorders characterized primarily by a progressive and severe weakness and muscle tightness (spasticity) in the lower limbs [1], [2]. HSPs are usually divided into pure and complex forms, the latter being characterized by the presence of additional symptoms, including neuropathy, parkinsonism and cognitive impairment [1], [2]. Advances in sequencing technologies have revealed that HSPs are among the most genetically varied disorders, with mutations at more than 80 genetic loci and different possible transmission modes, such as X-linked, maternal linked, autosomal-dominant and autosomal-recessive. Due to these mutations, aberrant proteins are expressed leading to the degeneration of corticospinal axons controlling lower motor neurons [1], [2]. However, especially in pure forms, neuronal death is typically low. The functions of many proteins involved in the pathogenesis of HSPs are related to intracellular trafficking, biogenesis and/or distribution of membrane compartments, the regulation of signalling pathways important for axon homeostasis, shaping and positioning of organelles and signalling complexes (motor proteins), and axon myelination. Interestingly, only few HSPs proteins are associated directly with mitochondrial functions, whose impairment is an hallmark of neurodegenerative diseases [1], [2].

Infantile-onset ascending hereditary spastic paralysis (IAHSP) is a pure form of HSP inherited in an autosomal recessive manner [3]. It is characterized by the onset of spasticity to lower limbs within the second year of life, involvement of the upper limbs by the seventh to eighth year of life, wheelchair dependence starting from the second decade of life and progression towards pseudobulbar syndrome and spastic tetraparesis. Finally, cognitive functions are preserved [3]. This disorder is associated with mutations at the Amyotrophic Lateral Sclerosis type 2 (ALS2) gene, which encodes for Alsin protein, composed by 1657 amino acids organized in multiple domains [4]. Evidences revealed that Alsin plays a fundamental role in vesicular trafficking and neuronal homeostasis by means of its ability to interact with two guanosine triphosphatases, Rac1 and Rab5 [5]–[7]. In particular, *in vitro* studies suggest that its central region, composed by a Dbl Homology (DH) domain followed by a Pleckstrin Homology (PH) domain, can bind to Rac1. Moreover, this interaction has been demonstrated to be necessary for the following activation of Rab5 by the C-terminal region of Alsin [4], [7]. However, as far as we know, the physiological structure of this protein and its relationship with specific functions are still unknown.

Computational techniques have exponentially grown by virtue of remarkable improvements in computer hardware and software. In particular, Molecular Modelling is becoming an elective tool to study nanoscale level biological systems. Indeed, it is currently used to model the 3D structure of proteins starting from their amino acid sequence and to study their dynamics [8], [9]. With the aim of designing potential therapies for IAHSP disease, a proper understanding of Alsin biological functions at the molecular level is crucial. Therefore, the aim of this M.Sc. thesis is to develop a 3D model of Alsin DH/PH domain and study the effect of Rac1 interaction on its dynamics in order to relate biological and molecular evidences.

The work is organized as follows:

Chapter 1 is the present introduction, where the framework and the aim of the thesis are presented.

Chapter 2 provides a biological background about IAHSP disease, Alsin role in vesicular trafficking, and the effect of mutations on neurons. To a better understanding of Alsin biological functions, a paragraph will be dedicated to the description of guanosine triphosphatases and related proteins. Finally, a general description of DH/PH domains, their functions in different proteins, and previous computational studies on them will be provided.

Chapter 3 is a description of the methods employed in the present work. Molecular Modelling with a theoretical description of Molecular mechanics, Homology Modelling, and Molecular Dynamics. Then Markov State Models and Transition Path Theory will be presented.

Chapter 4 is the replica of previous results from literature on the DH/PH domain of a different protein. The work presented in this chapter will provide a validated experimental setup to be exploited in the analysis of Alsin.

Chapter 5 focuses on the construction of Alsin homology model and the analysis of its dynamics, both alone and in presence of Rac1. The effect of the interaction with the GTPase on DH/PH domain mechanical properties and conformations will be discussed, focusing on the relationship between the results and the evidence from previous literature.

Chapter 6 summarizes the results from chapter 4 and 5 and presents the future perspectives of this work.

2 Biological background

2.1 IAHSP: from macroscopic features to molecules

IAHSP is a rare autosomal recessive neurodegenerative disorder associated with mutations in the ALS2 gene, locus 2q33.1, which encodes for Alsin protein [6], [10], [11]. It is a pure form of HSP in which clinical presentation is rather homogeneous, with the onset of spasticity to the lower limbs within the second year of life. While some children are initially able to walk independently and then lose their ability, others never learn how to walk. The disease rapidly progresses leading to upper limbs involvement, severe spastic tetraparesis, and pseudobulbar syndrome, a condition in which the patient, unable to control facial muscles, has speaking and swallowing difficulties; sometimes, the pseudobulbar syndrome is also characterized by sudden and uncontrollable episodes of laughing or crying. Wheelchair dependence usually occurs during the second decade of life and some patients lose bladder and sphincter control in advanced stages, but cognitive functions are preserved and long-term survival is compatible with the disease [3], [10]. Magnetic resonance images are normal in children, however, older patients are characterized by cortical atrophy in motor areas and T_2-weighted punctuate hyperintensities in the corticospinal pathways of the posterior arms of the internal capsule and brain stem. Moreover, FLAIR-weighted or T_2-weighted hyperintensities of periventricular areas and spinal cervical atrophy, typical of other HSPs, are common. Nerve conduction velocities are normal and there is no sign of denervation, as shown by electromyography, but motor evoked potentials reveal severe impairments of the corticospinal tract, consistent with the degeneration of upper motor neurons. Somatosensory evoked potentials are normal in children, but not in older patients [3], [10]. It was observed that families with identical homozygous ALS2 variants demonstrate phenotypic variability, both intra- and interfamilial, suggesting that environmental and epigenetic factors may play a role in the disease [10]. The diagnosis is done through the identification of biallelic pathogenic variants in ALS2 gene on molecular genetic testing, that include single-gene testing and multigene panels including both ALS2 and other genes of interest.

The ALS2 gene, composed by 34 exons, is located on the long arm of chromosome 2q33. Alternative splicing of its transcripts generates two variants, a short form of 396 amino acids and a long form of 1657 amino acids, ubiquitously expressed in human tissues, especially in the spinal cord and the brain [12]. The long-form of Alsin is a 184-kDa protein composed of five main structured domains: starting from the N-terminus, there is a regulator of chromosome

condensation 1 (RCC1)-like domain (RLD), a Dbl homology (DH) domain, a pleckstrin homology (PH) domain, eight consecutive Membrane Occupation and Recognition Nexus (MORN) motifs and, finally, a vacuolar protein sorting 9 (VPS9) domain [4]. The short form of Alsin only encodes for a part of the RLD (Figure 1). Mutations can occur in different protein domains and be of a different type, such as frameshift, missense or nonsense, leading to the expression of different types of aberrant proteins. However, the analysis of such mutations and the resulting phenotypes could not find a direct correlation between symptoms and mutation site and/or type [13]. It is, therefore, necessary to investigate the molecular functions of this protein to understand how mutations could interfere with them, leading to upper motor neurons degeneration.

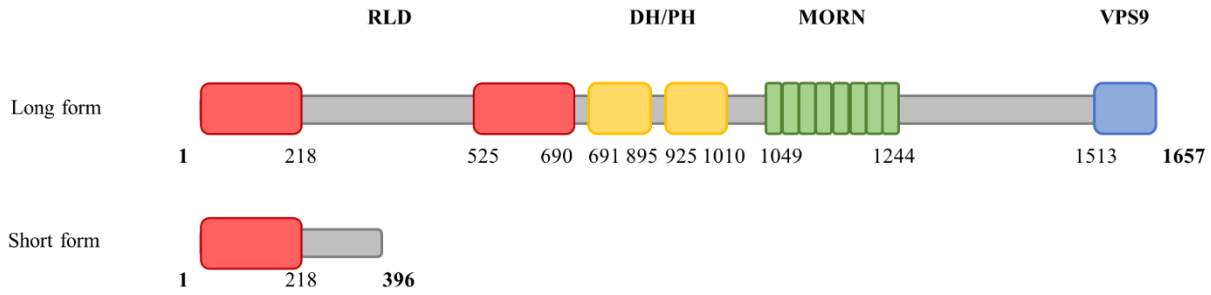

Figure 1. Schematic representation of Alsin domains in both the long and the short form.

2.2 GTPases and related proteins

To better understand the biological functions of Alsin, an introduction to guanosine triphosphatases (GTPases) is necessary. They are proteins capable of hydrolyse guanosine triphosphate (GTP) and with a high affinity for both guanosine diphosphate (GDP) and GTP [14]. They act as molecular switches, cycling between an inactive GDP-bound state and an active GTP-bound state. GTPases are classified based on functional, structural, and sequence similarity between them, but a first distinction can be made between trimeric and monomeric (also referred as small) GTPases. In particular, Ras superfamily of small GTPases can be divided into five branches [14]:

- ADPribosylation factor (Arf),
- Ras sarcoma (Ras),
- Ras homologous (Rho),
- Ras-like nuclear (Ran),

- Ras-like proteins in brain (Rab).

Arf proteins are involved in vesicular transport regulation, while Ran GTPases are important in the nucleocytoplasmic transport of both proteins and RNA. Ras family of proteins is known for its role in oncogenesis, since they activate in response to various extracellular stimuli and consequently control signalling networks that regulate cell differentiation, proliferation, and survival. Rho GTPases are known as key regulators of actin cytoskeleton reorganization in response to extracellular stimuli. As a consequence, these proteins have been associated to cell-cell and cell-matrix interactions, regulation of endocytosis and exocytosis, cell movement, and cell-shape. The most studied members of this family, composed by approximately twenty members, are Cdc42, RhoA, and Rac1. Finally, Rab family is the largest group of proteins in Ras superfamily. Its 61 members are involved in the regulation of intracellular vesicular transport, as they facilitate vesicle formation, fusion, release, and transport to the acceptor compartment. These proteins are located in specific intracellular compartments depending on their functions in different vesicular transport processes; for instance, Rab5 is sited in early endosomes and regulates the transport from the plasma membrane to early endosomes of clathrin-coated vesicles [14]. The main branches of Ras superfamily and their functions are summarized in Table 1.

Table 1. Summary of Ras superfamiliy groups and their main functions

Family	**Functions**	**N° genes**	**Examples**
Arf	Vesicular transport regulation	27	Arf1, Arf6
Ras	Cell differentiation and proliferation	36	Ras, Rheb
Rho	Actin cytoskeleton reorganization	20	Rac1, RhoA
Ran	Nuclear transport	1	Ran
Rab	Intracellular vesicular transport	61	Rab5, Rab7

Given the high affinity towards both GDP and GTP, the low GDP/GTP exchange activity, and the low GTP hydrolysis rate, the GTPase cycle is regulated by guanine-nucleotide exchange factors (GEFs) and GTPase-activating proteins (GAPs). A GEF protein is a positive regulatory protein that speeds up the intrinsic GDP/GTP exchange activity of GTPases, although this is not an active process since GEFs only facilitate the replacement nucleotides. Two main factors

favours the GTP-bound state: first, the much higher cytosolic concentration of GTP than GDP and, second, the fact that GTPases in the active GTP-bound state have lower affinity towards GEF proteins than in the inactive GDP-bound state. Therefore, the exchange of GDP with GTP is more probable [14], [15]. However, the low hydrolysis rate of GTPases not only limits the efficiency of signal transduction but maintains for an excessive time the active state of these proteins. For this reason, GAPs stabilize the hydrolytic machinery of GTPases, accelerating GTP hydrolysis and making signalling processes more dynamics and efficient. Both GEFs and GAPs can be multidomain proteins with specific regulatory sites for protein or lipid interactions. For instance, phosphatidylinositol phosphates (PIPs) are a family of phospholipids present in cell membrane involved in the recruitment of different proteins at this level [14], [15]. Finally, effectors are proteins with higher affinity for GTP-bound GTPases that transduce their signalling to achieve specific biological functions and cellular response (Figure 2).

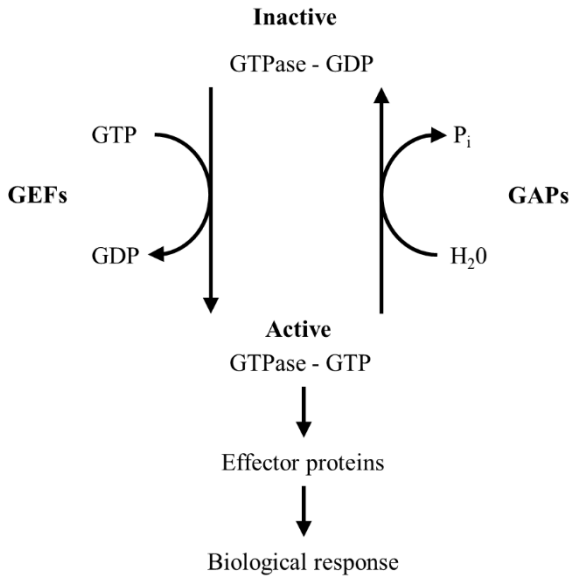

Figure 2. Representation of the GTPase cycle. An inactive GDP-bound GTPase can change its functional state thanks to the action of a GEF, which facilitates the release of GDP and, therefore, its exchange with GTP. On the other side, GTPases can be inactivated by a GAP, which accelerates its hydrolytic activity. The transduction of a signal carried by an active GTPase to effector protein leads to the activation of specific pathways and biological responses.

All GTPases are characterized by a G domain containing GTP-binding pocket and the hydrolytic machinery. In particular, two conserved loop-like regions, called switch I and switch II, are characterized by conformational changes between GDP- and GTP-bound state (Figure 3). Moreover, they are involved in interactions with binding partners, so that sequence differences are related to the selectivity of small GTPases [15].

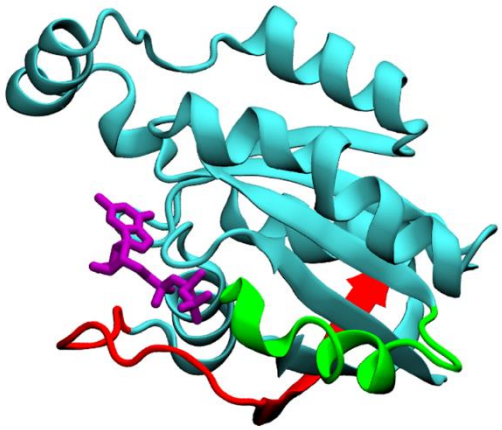

Figure 3. Crystal structure of GTP-bound Rac1 (PDB: 3TH5). Rac1 and GTP are coloured in cyan and purple, respectively. Switch I and II are highlighted in red and green [16].

2.3 Alsin and vesicular trafficking

Alsin contains three putative GEF domains: the N-terminal RLD, the central DH/PH domain and the C-terminal VPS9. The former is homologous to the protein RCC1, which acts as a GEF for Ran. However, Alsin has shown no GEF activity towards Ran suggesting, together with its cytoplasmic localization, that this domain has different functions [17]. The DH/PH motif is a common feature of different proteins known as GEFs for Rho GTPases. Therefore, the enhanced levels of active Rac1 in cells after the overexpression of ALS2 gene have suggested that Alsin acts itself as a GEF towards Rac1 [18]. Subsequent in vitro and in vivo studies have demonstrated that Alsin DH/PH interacts specifically with Rac1, but this interaction is not sufficient to activate it. Thus, it has been proposed that Alsin can act as a Rac1 effector rather than a Rac1 GEF [7]. Finally, the VPS9 domain was found to act as a GEF towards Rab5, leading to the enlargement of early endosomes. These findings are consistent with the subcellular localization of Alsin long form, which is located both in the cytosol and at early endosomes-level [5], [7], [18].

Alsin functions are related to the modulation of macropinosomes and early endosomes fusion and trafficking, especially in neurons [5]. An endosome is a vesicular structure involved in intracellular sorting and degradative pathways. It is possible to identify three main types of endosomes: early endosomes, late endosomes and recycling endosomes [19]. Focusing on the first two of them, early endosomes are tubular structures, usually colocalized with Rab5, that represent the first vesicles to fuse with endocytic material. They then mature to late endosomes, enlarged spherical structures usually associated with Rab7. Once these vesicles dissociate from Rab5 and Rab7, they can fuse with lysosomes leading to the degradation of their content [20].

Macropinosomes are different endocytic structures, involved only in degradative pathways, that forms due to complex signalling involving membrane ruffling, a Rac1-regulated process. In particular, it consists in the formation of membrane protrusions enriched with newly polymerized actin filaments. When the membrane protrusion collapses back, a large vesicle, i.e. a macropinosome, forms and then fuse with early/late endosomes [21]. Furthermore, Alsin also colocalizes with autophagosomes and/or hybrid structures formed by the fusion of autophagosomes and endosomes, i.e. amphisomes. Indeed, it has been demonstrated that Alsin overexpression enhances amphisomes formation through the activation of Rab5 [22]. As macropinosomes, autophagosomes are vesicles designated to degradative pathways, but they are part of the intracellular degradative system. In particular, in neurons all these vesicles need to be backpropagated to the cell body in order to fuse with lysosomes and break down their content. A schematic representation of this process is shown in Figure 4.

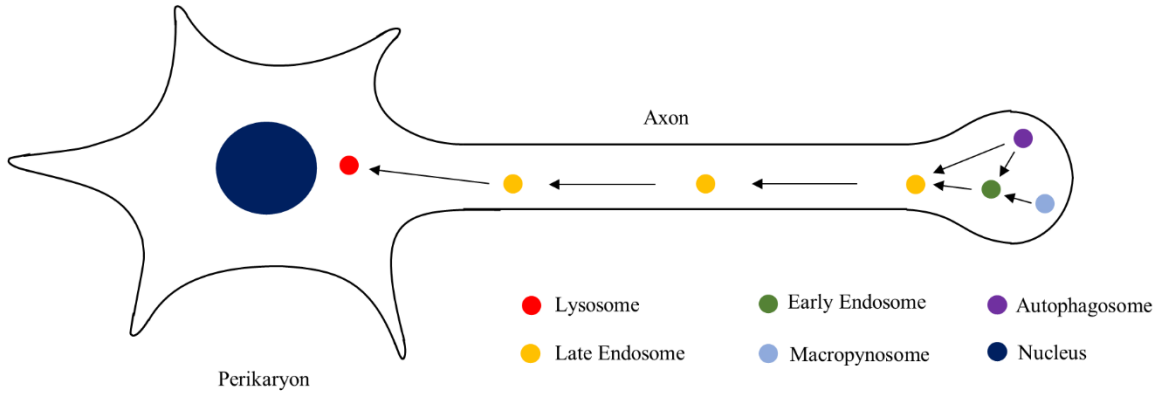

Figure 4. Vesicular trafficking and retrograde transport in neurons. Alsin biological functions are related to vesicular trafficking, especially in neurons where impairments of autophagy-endolysosomal system is associated with disturbed axon homeostasis and neurodegeneration. Indeed, degradation of vesicular cargos in lysosomes require the maturation of late endosomes, their backpropagation to the perikaryon and fusion to lysosomes. Alsin plays a crucial role in the formation of early endosomes, their fusion with autophagosomes, and their enlargement.

Of note, impairments the autophagy-endolysosomal pathways are associated with different neurodegenerative diseases since they disturb neuronal homeostasis and may lead to the accumulation of misfolded and/or aggregated proteins [23]. Moreover, macropinosomes trafficking is deemed crucial for axon outgrowth, process that requires a continuous supply of plasma and recycling of membrane proteins. In particular, the activation of Rac1-dependent pathways have demonstrated to facilitate axon outgrowth [7]. Hence, Rac1-induced activation of Alsin may be linked to neuron development and homeostasis [24].

2.4 Alsin and neuron degeneration

2.4.1 Tetramerization and subcellular localization

The C-terminal region of Alsin, spanning the MORN motifs and VPS9 domain, has shown the ability to dimerize in an antiparallel fashion and that this self-interaction, mediated by a.a. 1280-1335, is crucial for Rab5 activation *in vivo*, although not essential for Rab5 binding [25]. Moreover, an interaction between the N-terminal RLD and C-terminal region of Alsin (residues 1018-1657) is responsible for its sequestering in cytoplasm: the expression of proteins lacking RLD showed an increased endosomal localization of Alsin. However, deletion of RLD is not sufficient to fully relocalize Alsin from cytoplasm to early endosomes, suggesting that this translocation is triggered by an upstream activator. In particular, *in vitro* studies have demonstrated that Rac1 alters the subcellular localization of Alsin through its interaction with the DH/PH domain [7]. In a recent study, it has been observed that Alsin is able to further interact through the DH/PH domain, forming in this way a tetrameric complex necessary for its relocalization from the cytoplasm to the membrane and early endosomes. Since other DH domains are known to form homophilic dimers, it seems reasonable that this interaction is mediated by the sole DH domain (see section 2.5). On the other side, RLD is not necessary to form a proper oligomeric complex although retaining the ability to interact with Alsin C-terminal region [4].

Therefore, in absence of stimuli, Alsin is normally sequestered in cytoplasm due to a self-interaction between the RLD and C-terminal region. When Rac1 is activated by an upstream GEF, it binds to the DH/PH domain leading to a conformational change, tetramerization through DH domain and a.a. 1280-1335 (Figure 5), and relocalization to membrane ruffles. After Rac1 signalling, RLD is thought to have a role in ruffle localization due to its affinity to different PIPs, especially phosphatidylinositol-3-phosphate [PI(3)P] which is an important signalling lipid with a role in macropinosomes maturation and trafficking. At ruffles level, Alsin acts as a Rab5 GEF, inducing the fusion of the newly-formed macropinosomes with endosomes, their enlargement and maturation [4], [7], [26].

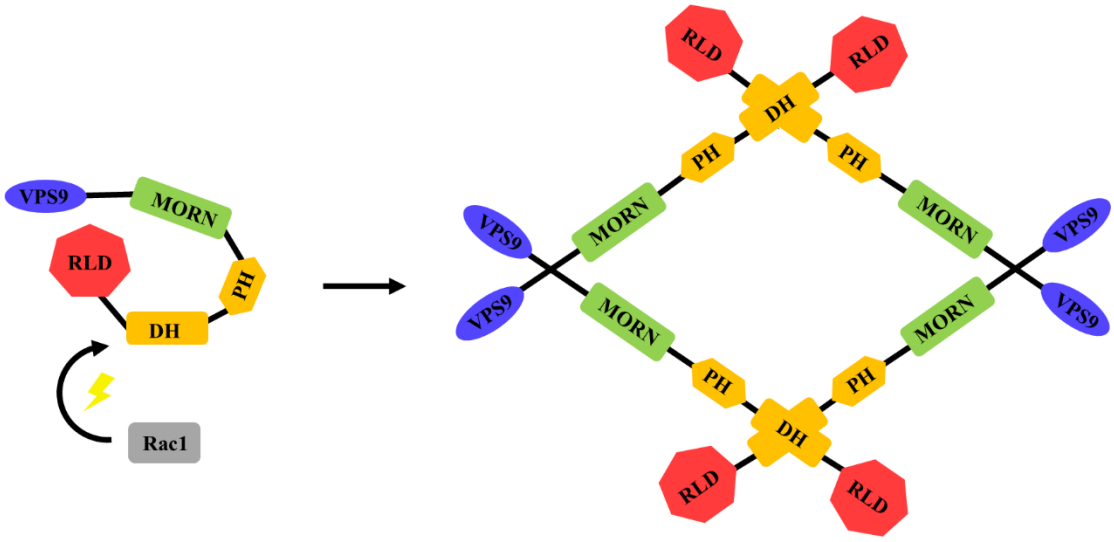

Figure 5. Tetramerization of Alsin due to Rac1 signalling. Alsin is thought to be sequestered in the cytoplasm in a closed form due to the interaction of its N-terminus and C-terminus. The interaction with Rac1 triggers a conformational change and the formation of an active tetramer, due to the dimerization through DH domains and C-terminal regions.

Alsin tetramerization is crucial for a proper relocalization of the protein to membrane ruffles and, consequently, Rab5 activation. Indeed, missense mutations in VPS9 domain correspond to lower molecular weight complexes, while mutations in RLD and PH domain lead to higher molecular weight structures All these aberrations do not allow a proper transition from the cytoplasm to cell membrane, although there is evidence that altered RLDs do not show decreased affinity to PI(3)P [4]. Moreover, evolutionary conserved residues associated with GEF activity of DH domain are not essential for Alsin transition to cell membrane [7]. This evidence is in agreement with Alsin being a Rac1 effector rather than GEF. However, it is not known whether missense mutations in the region responsible for dimerization (see section 2.5) inhibit ruffle relocalization.

2.4.2 Effects of Alsin loss in neurons

Several studies were made to investigate the effect of Alsin absence in mice, but no clinical manifestation of ALS2-related pathologies was obtained. However, significant alterations were observed in neurons derived from Alsin-knockout (KO) mice. First, axonal growth in hippocampal neurons was slowed demonstrating the crucial role of this protein in the neuronal development [24]. Detailed analysis of corticospinal motor neurons (CSMN) revealed signs of axonal degenerations, such as presence of membranous debris and collapsed synaptic vesicles. Moreover, an increased presence of autophagocytic vesicles containing broken mitochondria was detected selectively inside the apical dendrites of CSMN, but there was no significant loss

of these cells during time. This may represent a major problem since most of CSMN cortical inputs converge at apical dendrite level. As for mitochondria, major defects were present in the Golgi apparatus, which is fundamental in the control of secretory pathway, the post-translational modification of proteins, and intracellular vesicular transport [12]. Thus, loss of Alsin seems to affect selectively neurons of the corticospinal tract inducing common signs of neuronal degeneration.

The interaction between endosomal machinery and mitochondria is known to be important for cell homeostasis, repair, and apoptosis. Recently, the recruitment of Rab5-positive early endosomes to mitochondria was observed under oxidative stress condition, inhibiting cytochrome c release and, hence, increasing cell viability. Interestingly, Alsin was required for Rab5 translocation to mitochondria and localized at this level only under stress-induced condition. Indeed, Rab5 recruitment to mitochondria was severely reduced in ALS2-KO spinal motor neurons, suggesting a protective role of Alsin against oxidative stress [27].

Interestingly, Alsin can selectively bind to mutant superoxide dismutase (SOD1) through its entire DH/PH domain, while the deletion of DH or PH domains suppresses this function [18]. Moreover, it was proved that overexpression of ALS2 in transgenic mice carrying mutation on SOD1 gene decreases ROS production, demonstrating Alsin role in pro-inflammatory signalling regulation [18], [28]. Protection against excessive oxidative stress is particularly important in non-divisible cells like neurons, for whom the damage is cumulative. Transgenic ALS2-KO mice were shown to be more sensitive to cell death due to oxidative stress, meaning that the loss of Alsin predisposes rodents to oxidative stress but is not enough to induce neuronal degeneration [29].

Together with augmented oxidative stress, protein aggregation and disfunctions in cell clearance pathways are common features of neurodegenerative diseases. It was observed that loss of Alsin prevents the fusion between endosomes and autophagosomes, obstructing the autophagy-dependent protein degradation [22]. Interestingly, a recent study showed the presence of granular misfolded SOD1 inclusion in motor neurons of the spinal cord in a patient with ALS2 mutation [30]. Rab5 activation is, indeed, a crucial step for the retrograde axonal transport of vesicles to the cell body, where they can fuse with lysosomes and proceed with the degradation of their cargos. This mechanism is essential for neuronal survival and its impairments may lead to increased susceptibility to neuronal defects [31].

2.5 The DH/PH domain

The DH/PH motifs, found in many proteins, are characteristics of the Dbl family of Rho GEF: in this structure, the DH domain is responsible for the GEF activity while the PH domain regulates the exchange factor activity and the interaction with plasma membrane and actin filaments of the cytoskeleton [32].

DH domains are usually composed of 11 α-helices folded into an elongated, flattened α-helix bundle; however, some of them can combine to form a bundle of six major α-helices (α1-6) [33]–[35]. Three conserved regions are known: conserved region 1 (CR1), conserved region 2 (CR2), and conserved region 3 (CR3), that corresponds to great part of α1, α2, and α5 helices, respectively. CR1 and CR3, together with a part of α6, constitute the Rho GTPase interacting pocket, located near the centre of one surface. CR2, instead, is exposed on the opposite surface. PH domains are small domains present in various proteins composed by two anti-parallel β-sheets followed by a C-terminal amphipathic helix (Figure 6). Since the loops connecting the β-strands may have different lengths, it can be difficult to identify this domain [33], [36]–[38].

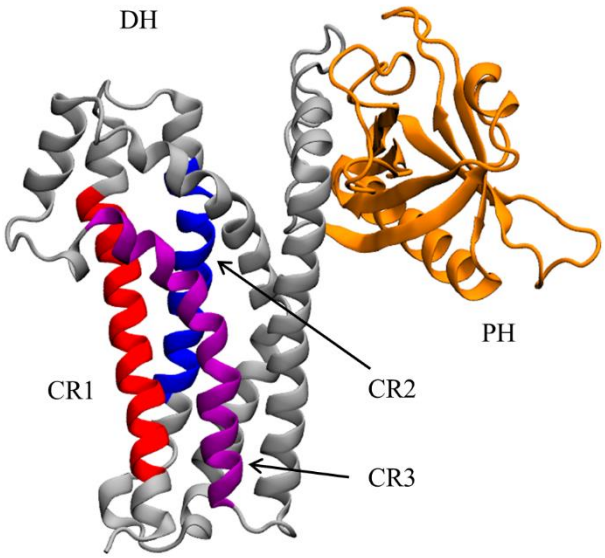

Figure 6. DH/PH domain of TIAM1 (PDB: 1FOE). The first, second, and third conserved regions in DH domain are represented in red, blue, and purple. PH domain is coloured in orange [39].

Alsin has been predicted to fold in a DH/PH like domain in its central region (residues 691-1010). As we have previously mentioned, the interaction of Rac1 with the DH/PH domain of Alsin has been identified as the trigger of oligomerization and relocalization of the protein. For this reason, here we focus on the description of this region, its main features and its functions on other proteins.

The DH domain of Alsin (residues 691-895) is composed by 206 amino acids while the PH domain by 86 (residues 925-1010).

Since Alsin DH/PH domain demonstrated not to act as a Rac1 GEF despite being predicted with the same structure of Dbl family, the functions of proteins containing one or both motifs have been investigated to understand the specific role of these regions. The DH domain is characteristic of proteins with known Rho GEF properties. Among them, ARHGEF6 and VAV2 play an important role in lamellipodia assembly, angiogenesis, and cell migration modulating Rho GTPases activity. On the other side ARHGEF10 has a role in myelination of peripheral nerves [40]. In protein ECT2, the PH domain folds back to inhibit the DH region catalytic activity thanks to the presence of a flexible disordered linker. The interaction of an active RhoA with the PH motif induces a conformational change and allows DH region to bind a second RhoA [41]. Notably, some proteins are characterized by the presence of two consecutive PH domains, as for FARP2. This Rac1 GEF is involved in neurite outgrowth and is allosterically auto inhibited by the PH motifs. After the phosphorylation-mediated activation, they move away to expose Rac1 binding surface [42]. The NET1 protein is a RhoA GEF that moves from the cytoplasm to cell membrane after Rac1 signalling. It seems that this process does not require neither the catalytic activity of DH domain nor the presence of PH region [43]. In particular, this process may be similar to the one following the interaction between Rac1 and Alsin. Hence, DH-containing proteins are associated with actin microfilaments dynamics, which is crucial in the formation of macropinosomes and endosomes, cell spreading and migration, neurite outgrowth, and lamellipodia formation. Although Alsin is not a GEF, as a downstream effector of Rac1 it is reasonable its involvement in similar processes.

As mentioned before, the presence of a PH domain is not a sole feature of Dbl family. In most proteins containing this motif, it is crucial for PIPs binding and membrane relocalization, allowing the protein to transfer a signal from the cytoplasm to the membrane. Typically, PIPs binding site is located near the loop between first two strands, which contains positively charged residues able to interact electrostatically with the negatively charged phosphoinositide head group [42]. In particular, TBC1 domain family member 2A is a Rac1 effector that moves to the membrane and activates Rab7. This transition is triggered by Rac1 binding to a coiled coil region near the PH domain [44]. The ACAP1 protein is involved in membrane remodelling and, to this purpose, dimerize through its BAR domain. In this way the PH domains are exposed and can interact with the membrane [45]. In a similar way, myotubularin-related protein 5 is able to form homodimers and eterodimers via interactions mediated by a coiled coil region that

precedes the PH domain [46]. The unconventional myosin-X is a protein containing an N-terminal motor region that is inhibited by the C-terminal PH/FERM domains. Phosphatidylinositol (3,4,5)-trisphosphate binding to PH domain stops this inhibition and allows the formation of an active dimer, exposing the tail region responsible for the dimerization [47]. This mechanism of autoinhibition of one active domain by the other terminal of the protein may be similar in Alsin. Indeed, it has been proposed that Rac1 activation may induce a conformational change of the protein, moving RLD away from VPS9, and allowing tetramerization [4]. Another example of autoinhibition is that of Rac-alpha serine/threonine-protein kinase: it is inactive due to intramolecular interactions of PH and kinase domains, but the interaction of the former with 3-phosphinositides induces a conformational change, moving PH region and allowing the activation of the protein [48]. Finally, Src kinase-associated phosphoprotein 2 is able to form a homodimer through the N-terminal region preceding PH domain; this dimerization, together with PH domain binding to PIPs, seems to control the functions of the protein [49].

Taken together, these information about DH and PH domains suggest that PH region is usually not involved in protein-protein interactions. However, it often plays a crucial role in the conformational changes associated to signalling processes. On the other hand, some Rho GEFs can form homophilic dimers through the sole DH domain and their dimerization is correlated with the GEF activity *in vivo*. In particular, the surface of Rho GEFs opposite to the one interested by the GTPase interaction is involved and specific biological essays demonstrated the CR2 to be crucial in the dimerization [36]. This evidence is consistent with the crystallographic structure showing the DH/PH domain of TIAM1 bound with Rac1. Here, the protein forms a dimer interacting through the surface opposite to Rac1 binding pocket (Figure 7) [39]. Moreover, the 3D structure of this dimer is compatible with the proposed tetrameric structure of Alsin [4]. Finally, it was demonstrated that dimerization of Dbl's big sister (Dbs), a protein characterized by the DH/PH motif and Rho GEF activity, is sufficient to trigger the membrane relocalization by coupling multiple PH domains [50].

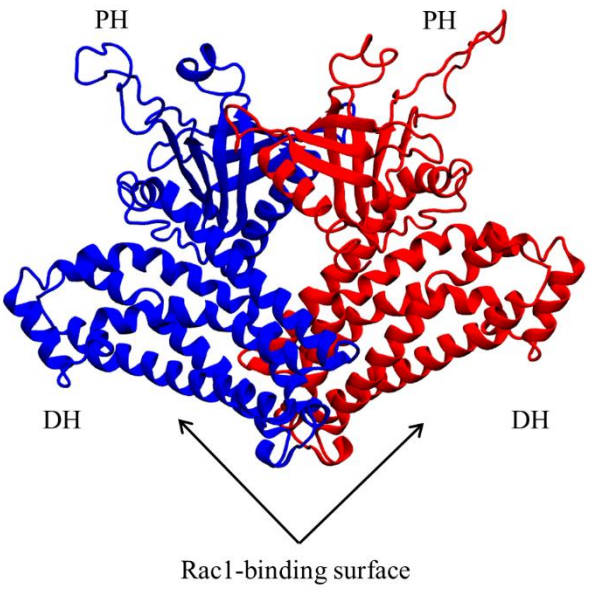

Figure 7. First and sixth chains of TIAM1 crystal structure (PDB: 1FOE [39]*). The conformation of the dimer obtained during the crystalization of the protein in presence of Rac1 is compatible with the arrangement of the probable Alsin tetramer.*

2.6 Computational modelling

It is now established that the functions of a protein largely depend on its 3D structure [51], therefore homology modelling is widely used to build atomistic models of proteins starting from their amino acid sequences. These informations can be exploited through Molecular Dynamics, which is emerging as an elective tool to study biological systems at the nanoscale level and relate macroscopic behaviour to its microscopic properties [8], [9].

Some studies have been carried out about the dynamics of the DH/PH domain in Rho GEF proteins. The analysis of several members of this family, both in their free form and bound to RhoA (a Rho GTPase), showed that the essential dynamics of the domain is characterized by a collective motion of the terminal part of α6 helix and the whole PH region, independently of the functional state. Consistently with these results, the same region expresses a great fraction of the overall flexibility in the DH/PH domain. Moreover, these studies highlighted the interaction of RhoA with CR1 and CR3 but also with PH domain, confirming its regulatory role in the catalytic activity of this family of proteins [34], [52].

3 Materials and Methods

The term Molecular Modelling refers to a set of tools that can be used to study the properties of molecular structures at a given condition by representing and simulating their behaviour [53]. The broad applications of Molecular Modelling can vary from material science to biology and involve different fields of expertise, such as physics, chemistry, engineering, and biology. Given the complexity and the high number of particles forming the systems of interest, an analytical treatment is almost impossible. Therefore, Molecular Modelling is now strictly related to numerical methods and computer modelling. In particular, macroscopic properties of complex systems, like biological ones, can be described ignoring the electrons behaviour and analysing only the motion of atoms nuclei, which is the approach of Molecular Mechanics.

3.1 Molecular Mechanics

Molecular Mechanics (MM) describes a system by means of the nuclei position, introducing a simplification that allowed to simulate systems of thousands of atoms. Indeed, this method is based on the Born-Oppenheimer approximation, which assumes that electron clouds instantaneously adjust after a change in the nuclear configuration since the electrons have a much lower mass than nuclei. Moreover, the latter are approximated as particles following the classical laws of Newton. Therefore, this approach is not able to represent events related to the description of electron density clouds, such as formation or disruption of covalent bonds. However, MM allows to consider a small number of cases to develop and test parameters, collected in what is known as force field, that can be used to describe the potential energy surface in a broad range of applications.

In MM, the potential energy surface, described as a function of the atomic positions, is the sum of two terms describing the intra- and inter-molecular forces contributions. Therefore, the potential energy $\mathcal{V}(r^N)$ can be written as a function of the position r of the N particles in the system:

$$\mathcal{V}(r^N) = \mathcal{V}_{bond}(r^N) + \mathcal{V}_{non-bond}(r^N) \quad (1)$$

The contribution given at the energy by intra-molecular interactions, also known as bond interactions, can be of three main types: bond stretching between two particles, angle bending between three particles, and bond torsion between four particles.

$$\mathcal{V}_{bond}(r^N) = \sum_{bonds} \mathcal{V}_{bonds}(r^N) + \sum_{angles} \mathcal{V}_{angles}(r^N) + \sum_{torsion} \mathcal{V}_{torsion}(r^N) \quad (2)$$

Different models have been proposed to describe each term contributing to the bond potential energy function. The most common and simplest way to model the energy associated to the stretching of a covalent bound between two atoms is Hook's law. This formulation depends on two parameters: k, the stretching constant of the bond expressed in kcal mol^{-1} Å^{-2}, and l_0, the reference bond length between the two particles interacting (Figure 8 A). Through this expression, a penalty to the potential energy functions is applied when bond length deviate from the reference l_0.

$$\mathcal{V}_{bonds}(l) = \frac{k_i}{2}\left(l_i - l_{i,0}\right)^2 \quad (3)$$

The Hook's law has also been used to model as a harmonic potential the angles term, which accounts for the energy associated to the valence angle. It is defined as the angle formed by three atoms i-j-k where i and k are both bonded to j (Figure 8 B). Therefore, two parameters are used to describe the angles term: h, the force constant expressed in kcal mol^{-1} deg^{-2}, and θ_0, the reference value of the valence angle. Compared to the energy required to stretch or compress a bond, lower energy is required to distort an angle.

$$\mathcal{V}_{angles}(\theta) = \frac{h_i}{2}\left(\theta_i - \theta_{i,0}\right)^2 \quad (4)$$

The torsional or dihedral term is defined between four atoms and refers to the rotation of a bond (Figure 8 C). It usually follows a sinusoidal law depending on three parameters: V_n is the barrier term and is related to the energy necessary to perform a rotation; n is the multiplicity and gives the number of minima encountered in a 360° rotation; γ is the phase factor and defines the position of the function minima.

$$\mathcal{V}_{torsion}(\phi) = \frac{V_n}{2}(1 + \cos(n\phi - \gamma)) \quad (5)$$

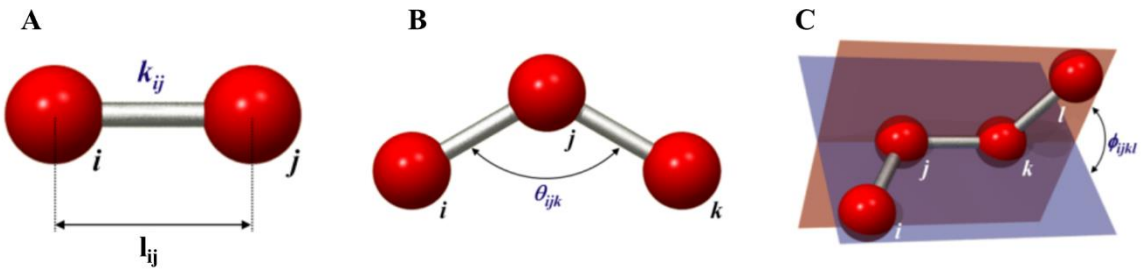

Figure 8. Bond interactions. (A) Bond length between atoms i and j. (B) Bond angle between atoms i, j, and k. (C) Torsion angle between atoms i, j, k, and l. Source: cbio.bmt.tue.nl/pumma/index.php/Theory/Potentials.

The inter-molecular or non-bond interactions are the way independent particles interact without having any specific bound or relationship between atoms. The energy associated to this kind of interactions is commonly proportional to a reverse power of the distance between atoms. Among the non-bond interactions, we can distinguish the electrostatic force and the van der Waals force. The former is related to the presence of unequal distribution of charges in the molecules due to the presence of species with different electronegativity. This phenomenon is usually modelled by the presence of different point charges, called partial or net atomic charges, localized at the nuclei centres. Then, the contribution of this type of interactions to the potential energy function is defined through the Coulomb law: the energy associated to two atoms with net atomic charges q_i and q_j at a distance r_{ij} is

$$\mathcal{V}_{electrostatic}(r_{ij}) = \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \tag{6}$$

where ε_0 is the dielectric constant. The van der Waals force is repulsive and grows exponentially at short distances, is attractive at longer distances and becomes nearly nil starting from few nanometres. The most common mathematical description of this function is the Lennard-Jones 12-6, which depends on two parameters specific for a pair of atom types interacting: the well depth ε and the collision diameter σ (Figure 9).

$$\mathcal{V}_{vdW}(r_{ij}) = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{7}$$

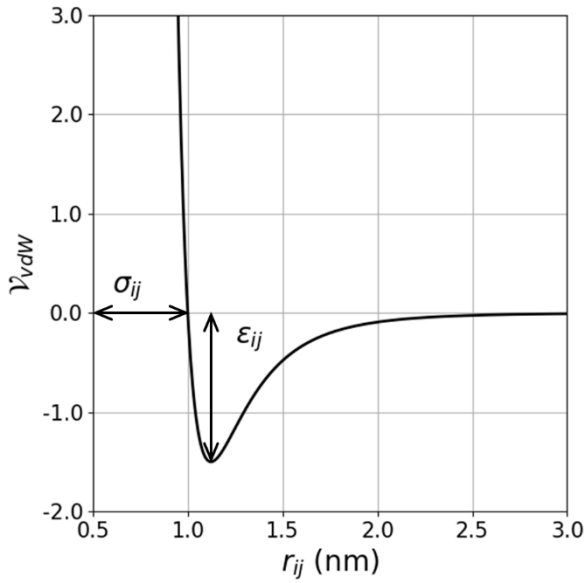

Figure 9. Potential energy of van der Waals interaction modelled with Lennard-Jones 12-6 potential. In this example, $\varepsilon_{ij} = 1.5$ *and* $\sigma_{ij} = 1$ *nm.*

Therefore, the inter-molecular contribution to the potential energy function can be expressed as

$$\mathcal{V}_{non-bond}(r^N) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \mathcal{V}_{electrostatic}(r_{ij}) + \sum_{i=1}^{N} \sum_{j=i+1}^{N} \mathcal{V}_{vdW}(r_{ij}) \qquad (8)$$

Non-bond interaction calculations are extremely expensive since their number grows with the second power of the number of particles N. Different methods have been proposed to reduce the computational effort, such as ignoring the non-bond interactions for particles whose distance is over a certain cut-off, potential switch, and particle mesh Ewald.

During a simulation only a finite number of particles can be considered, therefore the system is usually inserted in a box of different possible shapes (e.g. cubic, truncated octahedron, and dodecahedron). However, boundary conditions are crucial during the simulations because they strongly influence the properties of the entire system. To avoid edge effects, periodic boundary conditions are usually applied, i.e. the box is replicated in all directions so that particles on one side of the box see the periodic repetitions of particles of the other side of the box. Moreover, if a particle moves out of the simulation box, a replica enters from the other side to maintain constant the total number of particles inside the box. As a consequence, an upper limit to the cut-off value is necessary to avoid artifacts and respect the *minimum image convention*, i.e. to impede an interaction of a particle with itself (Figure 10).

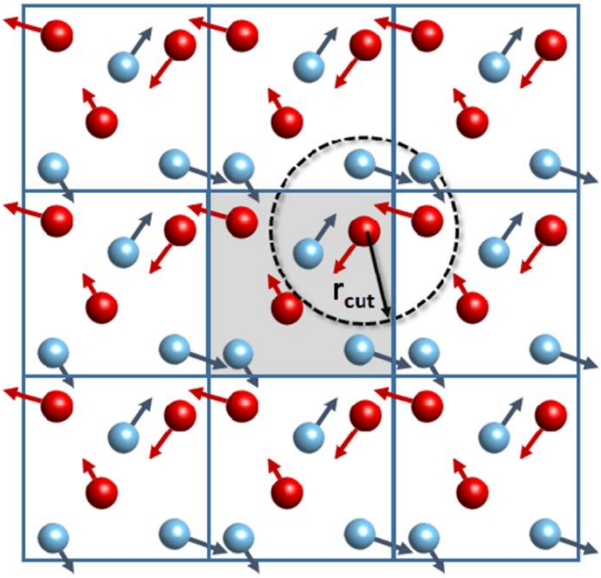

Figure 10. Representation of the periodic boundary conditions for a cubic box. The cut-off radius to respect the minimum image convention is showed [54].

The potential energy function describes a multidimensional surface of 3N cartesian coordinates, where N is the total number of particles in the system. The representation of this surface is almost impossible and feasible only for few cases. MM is interested in the minima of this surface since they correspond to stable arrangements of atoms. Therefore, potential energy minimization is a crucial step in Molecular Modelling, both as an integral part of techniques and as a way to prepare the system before other kinds of calculations, such as Molecular Dynamics. Indeed, it allows to avoid unstable interactions in the initial configuration of the systems that would produce high forces at the very beginning of a simulation. The different methods proposed to find a local minimum of the potential energy surface can be divided in derivative and non-derivative methods. Among the latter there are *simplex* and *sequential univariate* methods, while the former can be further classified in methods of the first order and of the second order. The first group is based on computing the gradient of the potential energy, since its direction is related to the position of the minimum. Some known methods of the first order are *steepest descend*, *conjugate gradients minimization*, and *line search in one dimension*. The methods of the second group exploit the information of the second derivative about the curvature of the potential energy function. Among them, there are *Newton-Raphson method* and *Broyden-Fletcher-Goldfarb-Shann method*. None of this method in always preferrable to the others, but the choice should be made according to the desired accuracy, storage capability, robustness, and the possibility to compute the second derivative.

3.2 Homology Modelling

A central issue in computational Molecular Modelling is obtaining reliable 3D all atom protein structures to be used as input in simulations. Despite the great number of experimental structures now available (e.g. in the Protein Data Bank [55]), techniques like X-ray crystallography and cryo-EM are time-consuming and expensive. Moreover, many protein structures have not been resolved yet or are difficult to obtain. To overcome these problems, Homology Modelling can be used to build atomic-resolution models of a protein (*target*) starting from its amino acid sequence and one or more experimental structures of homologous proteins (*templates*), i.e. proteins sharing a significant amino acid sequence identity with the target. Indeed, it is known that proteins with sufficiently high sequence identity (e.g. more than 40%) tend to have similar tertiary structures [56].

Given a protein to be modelled by homology, the first step is to identify homologous proteins with known 3D structures. Then, one or more of them should be selected as templates, usually choosing within the same family of the target and with the best possible resolution. Once the template has been identified, it is necessary to perform sequence alignment, i.e. a bioinformatics technique used to obtain the optimal alignment between two or more DNA, RNA, or protein sequences. To improve the correlation between computed sequence alignments and biological similarities, several algorithms and scoring schemes have been proposed. The latter are also crucial to define the similarity between two protein sequences, since a mismatch may have completely different biological outcome depending on the involved amino acids. For instance, a substitution of polar residue with a hydrophobic one may be more adverse than the substitution between to polar amino acids.

After the alignment, an initial model is built following the template and starting from the regions without gaps. While single amino acids deletions are acceptable to build the model, larger missing regions should be treated either using *ab initio* techniques, which try to model small sequences without any template, or searching a different template for the specific part. Once the folds corresponding to gaps are inserted, the model is refined using MM techniques to obtain a low energy conformation. The final model should be further analysed to assess is quality, such as investigating the presence of amino acids in the inaccessible regions of the Ramachandran plot. Moreover, Molecular Dynamics simulations are usually performed to fully equilibrate the obtained structure and analyse its stability.

Among many different tools for Homology Modelling, I-Tasser has been ranked among the top methods in the Critical Assessment of protein Structure Prediction (CASP) since 2006. Its methodology is composed by four main steps and exploits different tools (Figure 11). In the first stage of prediction, the target sequence is compared to a nonredundant sequence database to build a sequence profile and infer the secondary structure through PSIPRED. These data are used by LOMETS to search suitable templates in the PDB library. In the second stage, continuous fragments from template structures and unaligned regions built through *ab initio* modelling are combined together. The structural assembly is done through a modified replica-exchange Monte Carlo simulation with specific restraints, from which low energy conformations are obtained and clustered by SPICKER. Cluster centroids are obtained averaging the 3D coordinates of the cluster members. In the third stage, the fragment assembly simulation is performed again starting from the selected centroids to remove steric clashed and refine the structures. The decoys are clustered and the centroids are used by REMO to generate the final model. Finally, in the fourth stage the target model is compared to the matching proteins in PDB library to obtain information about model quality and structural similarities [57].

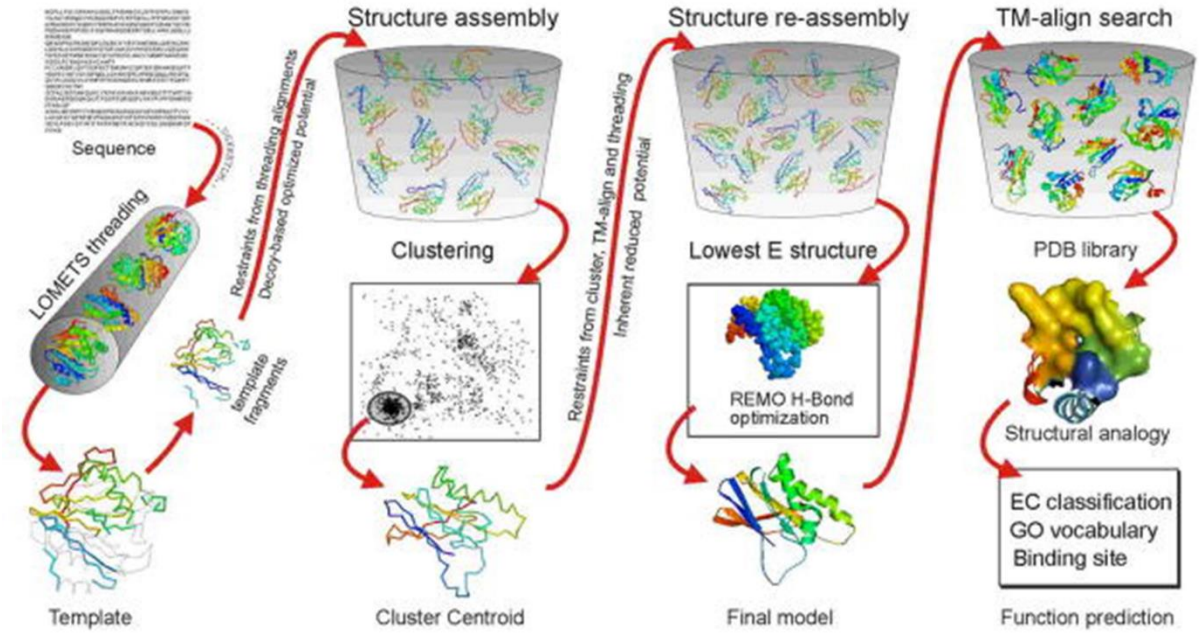

Figure 11. Schematic representation of I-Tasser methodology for protein structure prediction [57].

3.3 Molecular Dynamics

Despite providing tools to describe the mechanics of molecular systems and to find an energy minimum, MM and energy minimization can predict the properties only of quite simple systems. Indeed, only when all minima configurations are known it is possible to exploit statistical mechanics to obtain the thermodynamic properties of the system. Therefore, to obtain representative configurations of the system of interest other tools are needed, such as simulation techniques that can model its time-dependent behaviour. The most common simulation technique is Molecular Dynamics (MD), which solves Newton's equations for all the particles to obtain a trajectory in terms of positions and velocities.

The *phase space* of a system containing N atoms is a 6N dimensional space where a single configuration of the system is represented as a point defined by its 3N positions and 3N momenta. Every point in the phase space characterizes a microstate of the system, while a collection of microstates with the same macroscopic properties, e.g. temperature or pressure, is called a macrostate. A *statistical ensemble* is defined as a collection of points in the phase space that share the same macrostate. In MD the main statistical ensembles are:

- The micro-canonical ensemble (NVE), which describes an isolated system characterized by a fixed number of particles (N), assigned volume (V) and constant energy (E);
- The canonical ensemble (NVT), which describes a closed system with fixed number of particles (N), assigned volume (V), and coupled with a thermostat to maintain a constant temperature (T);
- The isothemal-isobaric ensemble (NPT), which describe a closed system with fixed number of particles (N) coupled both with a thermostat and a barostat to maintain constant temperature (T) and pressure (P);
- The grand-canonical ensemble (μVT), which describes an open system with fixed chemical potential (μ), volume (V), and temperature (T).

When analysing a system, one macroscopic property, that we can call A, is usually of interest. The value of A depends on the microstate of the system, therefore it is a function of the momenta $\mathbf{p}^N$ and positions $\mathbf{r}^N$ of all particles. The *ensemble average* of A can be obtained integrating over all possible microstates of the statistical ensemble.

$$\langle A \rangle = \iint A(\boldsymbol{p}^N, \boldsymbol{r}^N)\rho(\boldsymbol{p}^N, \boldsymbol{r}^N)\, d\boldsymbol{p}^N d\boldsymbol{r}^N \tag{9}$$

The function $\rho(\mathbf{p}^N,\mathbf{r}^N)$ is the probability density of the ensemble and its formulation depends on the type of statistical ensemble. In case of a canonical ensemble, defining with H the Hamiltonian, k_B the Boltzmann's constant, and Q the partition function, the probability density has the form of the Boltzmann distribution.

$$\rho(\boldsymbol{p}^N, \boldsymbol{r}^N) = \frac{1}{Q} e^{-\frac{H(\boldsymbol{p}^N, \boldsymbol{r}^N)}{k_B T}}; \quad Q = \frac{1}{N!} \iint e^{-\frac{H(\boldsymbol{p}^N, \boldsymbol{r}^N)}{k_B T}} d\boldsymbol{p}^N d\boldsymbol{r}^N \tag{10}$$

Hence, the partition function relates the microscopic state of a system with its macroscopic properties. The correct estimation of the ensemble average can be obtained only if all the possible states are known, but a complete sampling of the phase space is not feasible in a finite time. However, this problem can be overcome assuming the ergodic hypothesis, which states that an ensemble average can be replaced by a time average if the property of interest has been sampled for long enough.

$$\langle A \rangle_{ensembe} = \langle A \rangle_{time} = \lim_{\tau \to \infty} \frac{1}{\tau} \int_{t=0}^{\tau} A(\boldsymbol{p}^N(t), \boldsymbol{r}^N(t))\, dt \tag{11}$$

where $\mathbf{p}^N(t)$ and $\mathbf{r}^N(t)$ are the instantaneous momenta and position at time t, respectively. Since MD allows to obtain a trajectory of the systems in terms of M samples, numerical integration over a sufficiently long simulation, such that the phase space is correctly sampled, gives an estimation of the average of the property of interest.

$$\langle A \rangle_{time} \approx \frac{1}{M} \sum_{i=1}^{M} A(\boldsymbol{p}^N, \boldsymbol{r}^N) \tag{12}$$

In MD, the trajectory of particles is obtained in terms of positions and velocities solving Newton's equations of motion. This method is *deterministic* since, given an initial configuration of the system, each step depends only on the previous one. In particular, the acceleration *a* of each particle is obtained from the derivative of the potential energy function $\mathcal{V}$ with respect to its position *r*:

$$a = \frac{d^2 r}{dt^2} = -\frac{1}{m} \frac{d\mathcal{V}}{dr} \tag{13}$$

Due to the complexity of the potential energy surface, there is no analytical solution for this problem. Thus, the equations of motion are integrated using *finite difference method* and dividing the integration in small stages separated by a fixed time-step δt. The choice of this

parameter is crucial since too small time-steps increase the computational effort needed to explore a sufficient fraction of phase space, while too large time-steps may lead to instability and failure of the simulation. A suitable trade-off is represented by a δt ten times smaller than the fastest oscillation within the system, which is related to hydrogens in all atom simulations.

In a generic implementation scheme, initial atomic positions are derived from literature (e.g. Protein Data Bank) or from Homology Modelling, while starting velocities are assigned according to a Maxwell-Boltzmann distribution at a specified temperature. Before starting the simulation, temperature and pressure of the system are equilibrated while holding atomic positions fixed. Then, for as many steps as desired, the potential energy function is built exploiting the parameters of the selected force field and integration algorithms are used to obtain positions and velocities at the next step.

3.4 Principal Component Analysis

Due to the complexity of molecular systems, Molecular Dynamics trajectories are characterized by a large dimensionality. A reduction of degrees of freedom is often necessary to analyse simulation data and achieve greater interpretability. Principal Component Analysis (PCA) is a common technique used to dimensionality reduction purpose. It is a statistical procedure used to obtain from a set of variables a lower number of linearly uncorrelated features, called principal components (PCs), that are orthogonal direction of maximal variance.

The first step of this technique is the definition of the covariance matrix S of the atomic positions. The variance σ_v^2 of the data along a direction v can be written as:

$$\sigma_v^2 = v^T S v \tag{14}$$

Therefore, for Rayleigh variational representation, the direction v_1 of maximal variance σ_1^2 is defined as

$$\sigma_1^2 = \max_{|v_1|=1} v_1^T S v_1 \tag{15}$$

Thus, the first PC is the first eigenvector of the covariance matrix and the variance along it corresponds with the first eigenvalue. Through the diagonalization of the covariance matrix, 3N orthonormal eigenvectors (e_j), corresponding to the PCs, and the corresponding eigenvalues (σ_j^2) are obtained. The higher is the eigenvalue, the higher are the atomic fluctuations along the corresponding direction. Since the sum of all eigenvalues corresponds to the total variance of

the data, it is also possible to estimate the fraction of variance that is explained by the first m PCs, where usually $m \ll 3N$. In this way, it is possible to select a proper number of components to describe almost the totality of the trajectory of interest.

3.5 Markov State Models

While traditional MD is now an accepted tool to investigate molecular processes from a structural point of view and to relate them with experimental results, it is usually incapable of identifying different states and their kinetic relationship from the trajectories. One useful method to handle this problem is to build a Markov State Model (MSM), which consists in a network model of various states and the transfer constants between them [58]. In particular, if we consider a system in which n discrete states can be found, a MSM models the kinetics of the system through an $n \times n$ transition probability matrix estimated from MD trajectories. An essential feature of MSMs applied to MD is that they replace the view of single trajectories with an ensemble view of the dynamics. Moreover, trajectories used to estimate the model need to be long enough to reach only a local equilibrium rather than a global equilibrium, which may require orders of magnitude longer simulations [59].

3.5.1 Analysis of a continuous dynamics

To better understand the how a MSM is built from simulation data, it is necessary to describe first the ideal case of a continuous system. Consider a continuous state space Ω and a dynamical process $x(t)$; three main assumptions are made [59]:

1. $x(t)$ is a Markov process in the state space Ω, which means that the instantaneous change of the system depends only on $x(t)$ and does not require the knowledge of previous history. Consequently, the probability density to pass from $x \in \Omega$ at time t to $A \subseteq \Omega$ at time $t + \tau$ can be expressed as

$$p(x, A; \tau) = P\{x(t + \tau) \in A \mid x(t) = x\} = \int_{y \in A} p(x, y; \tau) dy \qquad (16)$$

2. $x(t)$ is ergodic, i.e. the state space does not have any dynamically disconnected states and all states will be visited for $t \to \infty$. The amount of time that the system spends in each state is given by the stationary density $\mu(x)$, which corresponds to the equilibrium probability density for a given thermodynamic ensemble.

3. $x(t)$ is reversible and, at the equilibrium, the number of transitions from *x* to *y* per time is equal to the number of transition from *y* to *x*. This assumption of *detailed balance* can be mathematically expressed as

$$\mu(x)\, p(x, y; \tau) = \mu(y)\, p(y, x; \tau) \tag{17}$$

This condition is a direct consequence of the second law of thermodynamics, since if it is not fulfilled there would be a set of states traversed in one direction with higher probability; if so, this preference of direction could be used to produce work from pure thermal energy.

If we consider an ensemble of molecular systems distributes in the state space at time *t*, its probability density $p_t(x)$ is different from the stationary density. If we wait a certain amount of time τ, this probability will change and become more similar to $\mu(x)$ since each system will undergo a transition according to the transition probability density $p(x, y; \tau)$. This phenomenon can be modelled through the propagator $\mathcal{Q}(\tau)$:

$$p_{t+\tau}(x) = \mathcal{Q}(\tau) \circ p_t(x) = \int_{y \in \Omega} p(y, x; \tau)\, p_t(y)\, dy \tag{18}$$

The same relationship can be described in the space of μ-weighted densities with the transfer operator $\mathcal{T}(\tau)$:

$$u_{t+\tau}(x) = \mathcal{T}(\tau) \circ u_t(x) = \frac{1}{\mu(x)} \int_{y \in \Omega} p(y, x; \tau) \mu(y)\, u_t(y)\, dy \tag{19}$$

Both operators fulfil the Chapman-Kolmogorov equation, i.e. they can be used to extend the dynamics for arbitrary long times. In case of the transfer operator, if $[\mathcal{T}(\tau)]^k$ is the *k*-fold application of $\mathcal{T}(\tau)$, the Chapman-Kolmogorov equation is

$$u_{t+k\tau}(x) = [\mathcal{T}(\tau)]^k \circ u_t(x) \tag{20}$$

Both the propagator $\mathcal{Q}$ and the transfer operator $\mathcal{T}$ are characterized by eigenfunctions and associated eigenvalues, that are contained in the interval [-1,1] if the dynamics is reversible.

$$\mathcal{Q}(\tau) \circ \phi_i(x) = \lambda_i \phi_i(x);\;\; \mathcal{T}(\tau) \circ \psi_i(x) = \lambda_i \psi_i(x);\;\; \phi_i(x) = \mu(x)\psi_i(x) \tag{21}$$

It is possible to observe a certain number *m* of dominant eigenvalues, associated to the first *m* slowest dynamical processes, while the remaining $\lambda_i < \lambda_m$ describe the fast processes that are usually not of interest. Thus, the dynamics can be described as the superposition of slow and

fast processes, where the latter tend to decay faster increasing the timescale of analysis [59]. In particular, there is one eigenvalue with the greatest norm $\lambda_1 = 1$ whose associated eigenfunctions are the stationary distribution $\mu(x)$ or a constant function, considering the propagator or the transfer operator, respectively. Regarding the remaining dominant eigenvalues, the closer they are to 1 the slower is the associated dynamical process, such that it is possible to obtain an implied timescale for each of them:

$$t_i = -\frac{\tau}{\ln \lambda_i} \tag{22}$$

Finally, the first *m* eigenfunctions of the transfer operator can be used to describe the transitions that characterize each dynamical processes according to their sign (Figure 12). For instance, in a system with four metastable states, if the *i*-th eigenfunction is positive in states A and B and negative in states C and D, it means that the *i*-th dynamical process is a transition between states A+B and C+D.

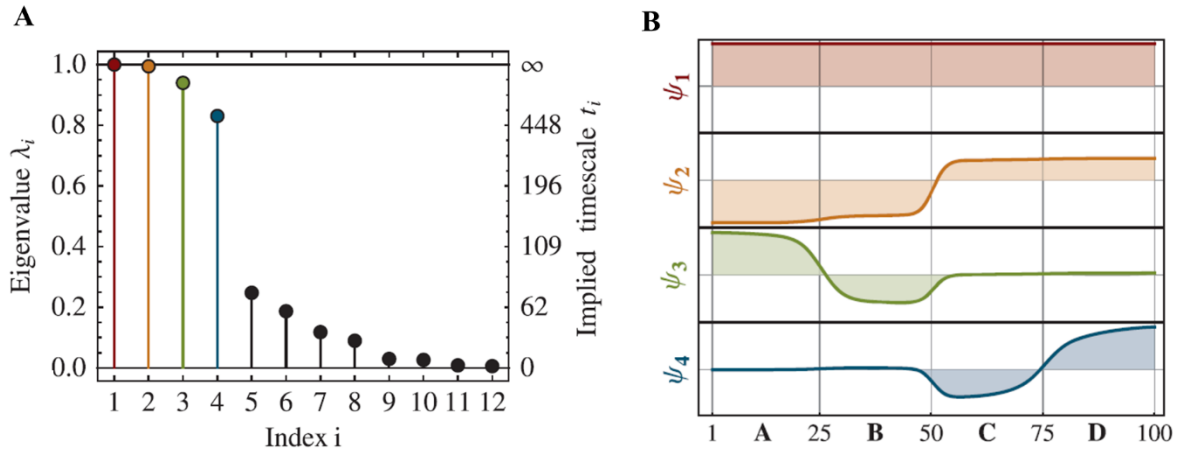

Figure 12. Eigenvalues and eigenfunctions of the transfer operator in an example system [59]. *(A) Representation of the eigenvalues. It is possible to observe a remarkable gap between the first four eigenvalues and the following ones, meaning that there are three dominant slow processes describing the system. (B) Representation of the right eigenfunctions corresponding to the slowest processes. Each eigenfunction describes the transition between the states in which it is positive and negative.*

3.5.2 Discretization of state space

In a real analysis, the state space is not continuous but should be discretized to obtain a computationally feasible description of the dynamics. Therefore, the transfer operator is approximated by a reversible transition matrix and the eigenfunctions $\phi(x)$ and $\psi(x)$ correspond to its left and right eigenvectors, respectively. A MSM then is a partitioning of the state space together with a transition matrix describing the jump processes between these discrete states in which the observed trajectories are projected. Consequently, the information

about where the original continuous process would be within a discrete space is lost and the jump process is no longer Markovian (Figure 13). The systematic discretization error that is introduced should be kept small enough to accurately describe the kinetics even for large and complex systems. Usually, the mentioned discretization is a simple partitioning with sharp boundaries in *n* states $S = \{S_1, S_2, \dots, S_n\}$.

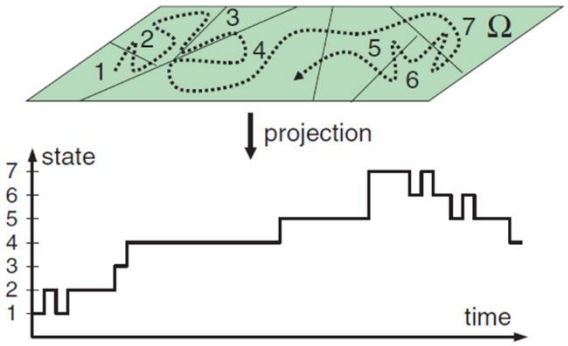

Figure 13. Example of discretization of the state space [59]. *The real continuous trajectory (dashed line) is projected onto discrete states, so that the result is a jump process between them.*

To represent the dynamics of the system, the degrees of freedom are usually reduced such as ignoring velocities or using specific coordinated defined by the analyst (featurization). After the discretization, the stationary probability to be in state *i* is given by the relation:

$$\pi_i = \int_{x \in S_i} \mu(x)\, dx \tag{23}$$

As mentioned before, the transfer operator is approximated by a matrix $\boldsymbol{T}(\tau) \in \mathbb{R}^{n \times n}$ such that the element $\boldsymbol{T}_{ij}(\tau)$ represents the probability for the system to be in state *j* at time $t + \tau$ given that at time t it was in state *i*, where τ is called the *lag time* of the model. An important feature is that to estimate the transition matrix the necessary dynamical information extends only over the lag time, i.e. no information about the global equilibrium is needed. If $\boldsymbol{p}(t)$ is a column vector containing the probabilities for the system to be in each of the *n* states at time *t*, we can obtain the same probabilities at time $t + \tau$ through the transition matrix:

$$\boldsymbol{p}^T(t + \tau) = \boldsymbol{p}^T(t)\, \boldsymbol{T}(\tau) \tag{24}$$

To model the system kinetics at time scales longer that τ, remembering that the transfer operator satisfies equation (20), we can use the expression:

$$\boldsymbol{p}^T(t + k\tau) \approx \boldsymbol{p}^T(t)\, \boldsymbol{T}^k(\tau) \tag{25}$$

This is only an approximation due to the discretization error, which depends on the discretization that has been chosen. In the analysis of molecular systems, the loss of information is due to both the discretization into a finite number of states and the reduction of coordinates used to describe the system itself. However, the finer is the discretization the smaller is the error, which should reduce increasing the lag time since we are less often imposing a local equilibrium into a discrete state [59].

3.5.3 Estimation and validation of the model

Suppose a trajectory generated at equilibrium conditions with N configurations stored every Δt and a discretization in *m* states. The information of the trajectory can be stored as the sequence of the discrete states $s_1, s_2, \dots, s_N$. It is important to underline that the trajectories can be of different length and restricted to different local equilibrium states, provided that the time resolution is the same. The first step is to choose a lag time τ, which should be an integer multiple of the data time resolution Δt. A common procedure to estimate the lag time is observing the value of the implied timescale of the slowest process at increasing lag times: a proper τ corresponds to the point where the curve reaches a plateau. Then, it is possible to define a count matrix $\boldsymbol{C}$ from the observed trajectories: the element $C_{ij}(\tau)$ is the number of times in which the system was in state *i* at time *t* and in state *j* at time $t + \tau$. In case of multiple trajectories, the single count matrices are simply added up. If $\chi_i(x_t)$ is the probability for the system to be in state *i* at time *t*, each element of the count matrix can be written as

$$C_{ij}(\tau) = C_{ij}(a\Delta t) = \sum_{k=1}^{N-a} \chi_i(x_k)\chi_j(x_{k+a}) \tag{26}$$

There are two main approaches to compute this matrix from data:

1. Subsampling of the trajectories at lag time to obtain statistically independent transition counts and a more robust estimation of the transition matrix. However, this approach may lead to a remarkable leakage of data and numerical problems;
2. Using a sliding window to count at lag time to avoid ignoring much of the data. However, nearby transitions, e.g. $t \to t + \tau$ and $t + \Delta t \to t + \Delta t + \tau$, are not statistically independent and the obtained model is only asymptotically correct. The bias introduced through this method also depends on the time resolution of the observed trajectory, since decreasing Δt the nearby transitions increase their correlation.

In case of an infinite long trajectory, the element T_{ij} of the transition matrix can be obtained by the number of *i*-to-*j* transitions divided by the total number of transitions from state *i*.

$$T_{ij}(\tau) = \frac{C_{ij}(\tau)}{C_{ij}(\tau)} \tag{27}$$

However, in the real case of finite trajectories a different approach to estimate the transition matrix which is more consistent with the observed data should be defined. To simplify the notations, in the following analysis the dependence on lag time of transition and count matrices is omitted. Supposing to have considered statistically independent transitions when building the count matrix, the probability that a transition matrix $\boldsymbol{T}$ would have generated the observed trajectories, i.e. the likelihood, is the product of the individual jump probabilities.

$$p(\boldsymbol{C}|\boldsymbol{T}) = \prod_{i,j=1}^{n} T_{ij}^{C_{ij}} \tag{28}$$

From the Bayes' theorem, the posterior probability of the transition matrix is proportional to the product of the likelihood and the prior probability of the transition matrix itself, before having observed any data:

$$p(\boldsymbol{T}|\boldsymbol{C}) \propto p(\boldsymbol{T})\, p(\boldsymbol{C}|\boldsymbol{T}) = p(\boldsymbol{T}) \prod_{i,j=1}^{n} T_{ij}^{C_{ij}} \tag{29}$$

The prior probability is crucial in the estimation of the transition matrix, therefore it should be chosen to obtain posterior distributions that lead to physically meaningful solutions. This is of particular importance when little observations are available, otherwise it should be sufficiently "weak" to give more emphasis on the observations. Usually, for computational simplicity, conjugate priors are usually chosen, i.e. functions that allow to obtain posteriors with the same functional form as the likelihood.

$$p(\boldsymbol{T}|\boldsymbol{C}) = \prod_{i,j=1}^{n} T_{ij}^{C_{ij}+C_{ij}^{prior}} = \prod_{i,j=1}^{n} T_{ij}^{C_{ij}^{tot}} \tag{30}$$

In the Bayesian approach, it is possible to estimate the transition matrix as the one that maximizes the posterior probability, which means using the so-called *maximum a posteriori estimator*. In this case, this approach leads to an asymptotically unbiased estimation of the transition matrix.

$$T_{ij} = \frac{C_{ij}^{tot}}{C_i^{tot}} \tag{31}$$

If no information about a prior distribution of the transition matrix is known, which corresponds to using a constant function as prior, maximizing the posterior is the same as maximizing the likelihood. This approach is named *maximum-likelihood estimation* and may lead to a significant bias in case of limited data. The obtained matrix expresses the transition probabilities between the *n* states. However, the number states chosen during the discretization is usually very high and much greater than the real number *m* of metastable states. For instance, two small states that are strongly connected and characterized by a high probability to observe a jump between them can be considered a unique state. To cluster together the initial states and found the *m* metastable states, algorithms like Perron Cluster Cluster Analysis (PCCA) can be used [58], [60].

Usually, we are also interested in the uncertainty of the estimation and, consequently, of the properties obtained from the transition matrix. In this case one commonly used method is the *Markov chain Monte Carlo sampling of transition matrices*, also known as Bayesian MSM. This approach consists in obtaining a certain number of matrices from the posterior distribution and use them to estimate also confidence intervals or standard deviations [61]. Notably, no assumption is made on the functional form of the posterior distribution. This uncertainty estimation is also important to validate the obtained model through the Chapman-Kolmogorov test. It basically tests whether the chosen discretization and lag time have led to a model that satisfies the approximation of the Chapman-Kolmogorov equation within statistical uncertainty:

$$[T(\tau)]^k \approx T(k\tau) \tag{32}$$

where $[T(\tau)]^k$ is the *k*-fold application of the transition matrix estimated from data and $T(k\tau)$ is the transition matrix estimated at a longer lag time $k\tau$ from the same data. Here, the statistical error due to limited sampling could be evaluated through a number of techniques, e.g. the Bayesian estimation [62]. In practice, the model to test is propagated at a longer timescale $k\tau$ through equation (24) and the obtained probabilities are compared with the elements of transition matrices computed at $k\tau$ in such a way to obtain a confidence interval: if the propagated probabilities lie in it, the test is passed. In a simplified version, only the self-transition probabilities are tested. In case of unsuccessful validation, the parameter used to build the MSM should be changed, e.g. the lag time. If this tuning does not solve the problems, new

coordinates should be used to describe the system or new MD data should be produced. The main steps involved in the estimation of an MSM are summarized in Figure 14.

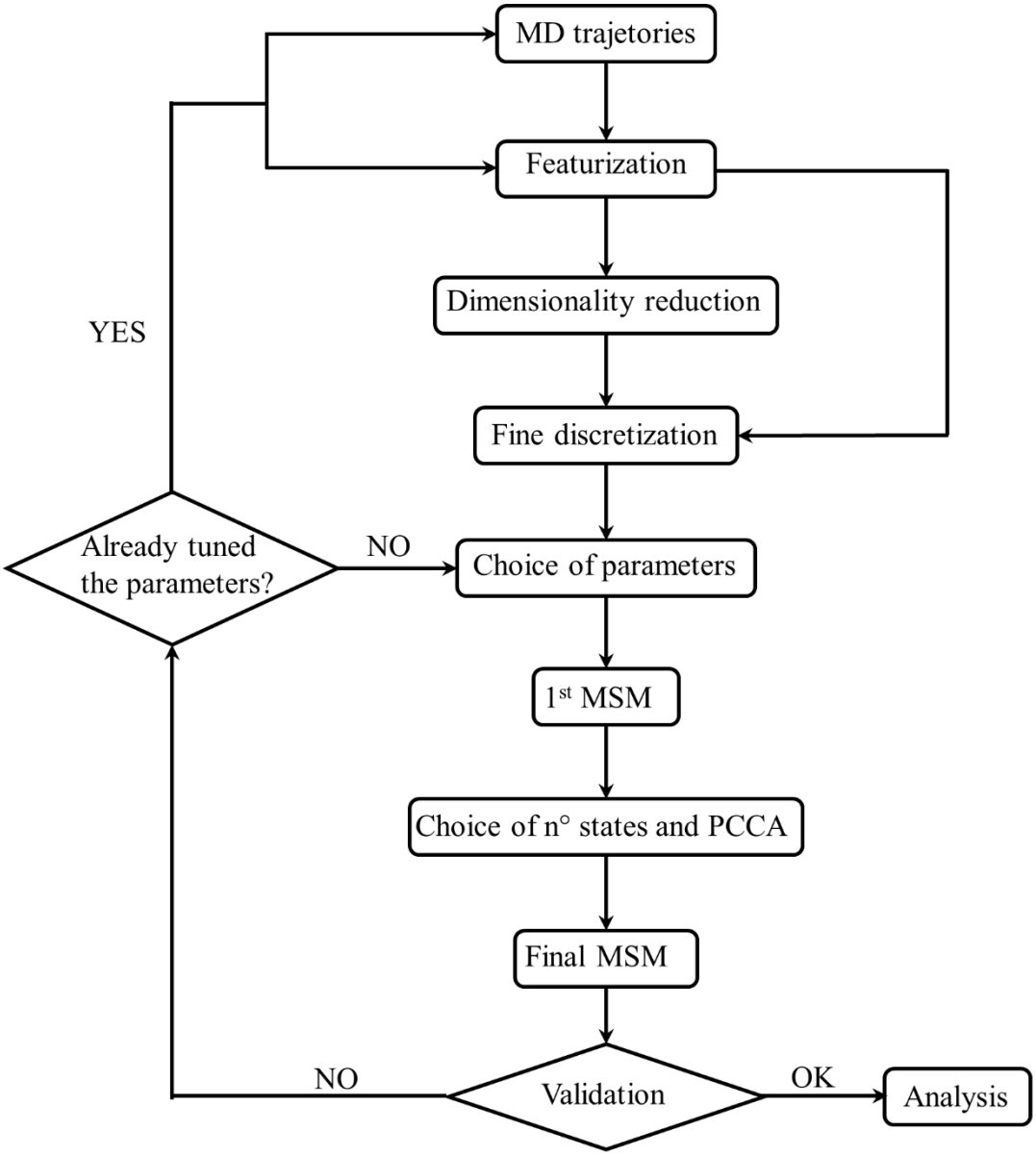

Figure 14. Flow chart representing the main steps involved in the estimation of an MSM from MD trajectories.

3.5.4 Transition Path Theory

In recent years, the continuous increase in the number of application of MSM has led to the need of new methods to analyse them, especially in case of large state spaces and complex networks. In this context, the framework of transition path theory (TPT) has been exploited to describe the statistical properties of transitions between the states defined by an MSM.

Basically, the idea is to consider two states of interest and find the typical mechanism by which the system jumps from the first to the second [63].

To better understand how is possible to find the most probable path connecting two states of a MSM, some initial definitions are necessary. If A is the starting state and B the ending state, a reaction event is an oscillations from A to B and the way this transition happens is of interest. In this framework, it is possible to introduce the concepts of *last-exit-before-entrance time* t^A, i.e. the time at which the systems last exit from state A before entering in state B, and *first-entrance-after-exit time* t^B, i.e. the time at which the system enters for the first time in state B on its way from state A. Therefore, the *n*-th reaction event starts at t_n^A, ends at t_n^B, and can be described by the ordered sequence of states visited during the transition from A to B. This latter is usually named as reactive trajectory. Finally, the distribution of reactive trajectories gives the probability that the system is at time *t* in state *i* and that is reactive:

$$m^R = \{m_i^R\}_{i \in S}; \qquad m_i^R = p(x(t) = i \ \cap \ t \in R) \tag{33}$$

where S in the state space and R is the set of reactive times $\{t_n^A, t_n^B\}$. Note that this quantity does not depend on time since *t* is fixed in the previous expression. Intuitively, this distribution can be thought as the product of the probability that, during the transition, the system arrived from A and the probability that it will reach B rather than A. To obtain a mathematical expression of this distribution, the concepts of foreward and backward committors have been introduced:

- The *i*-th element of the *forward committor* q^+ represents the probability that the process starting in state *i* will reach B rather than A;
- The *i*-th element of the *backward committor* q^- is the probability that the transition of system coming from state *i* started from A rather than from B.

It is possible to write those probabilities as:

$$q_i^+ = p(t_B^+ < t_A^+ \mid x(0) = i); \qquad q_i^- = p(t_B^- > t_A^- \mid \bar{x}(0) = i) \tag{34}$$

where t_A^+ and t_B^+ are the first time entering in state A and B, t_A^- and t_B^- are the last exit time from state A and B, x and $\bar{x}$ are the process and the time-reversed process. Therefore, ff π_i is the stationary probability for state *i*, m^R can be obtained through the following relationship [63]:

$$m_i^R = \pi_i q_i^+ q_i^-, \qquad i \in S \tag{35}$$

The committors are also useful to define an important quantity in transition path theory, called *probability current of reactive trajectories* f^{AB}, which is the average rate at which reactive trajectories flow from one state to another. Precisely, it is a matrix in which the element (i, j) is the limit for $s \to 0^+$ of the number N_s^{ij} of reactive trajectories jumping from state *i* to *j* in an interval long *s* divided by *s*. It is possible to prove that:

$$f_{ij}^{AB} = \lim_{s \to 0^+} \frac{N_s^{ij}}{s} = \begin{cases} \pi_i q_i^- T_{ij} q_j^+ & i \neq j \\ 0 & i = j \end{cases} \tag{36}$$

where T_{ij} is the transition matrix of the MSM.

Since the process is supposed to be reversible, transitions from state *i* to *j* and from state *j* to *i* can be observed at the same time. Therefore, one is usually interested in the *effective current* or *net flux* f^+, which is the net average number per time unit of reactive trajectories jumping from state *i* to *j* while moving from A to B.

$$f_{ij}^+ = \max\left(f_{ij}^{AB} - f_{ji}^{AB}, 0\right) \tag{37}$$

Now, it is possible to imagine modelling the system as described by the MSM with a graph, where the nodes represent the states and the edges are weighted by the elements of the net flux matrix (Figure 15). The graph contains multiple reactive pathways connecting the starting node A and the final node B. In each of them, the system has to move along edges associated with different effective currents. The edge with the minimum flux is referred as the bottleneck of the reaction pathway since it represents the slowest transition.

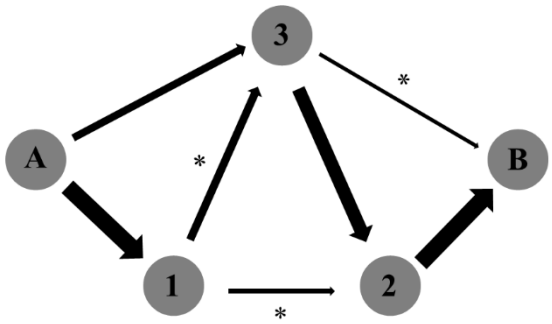

Figure 15. Graph representation of a five-state MSM. The starting and ending nodes of the transition are A and B, while states 1, 2, and 3 are intermediates. The width of the arrows represents the net flux between the states. Bottlenecks are marked with an asterisk. The most probable pathway, with the highest-flux bottleneck, is $A \to 1 \to 3 \to 2 \to B$.

Once we have built the graph, it is straightforward to define the best transition pathway as the one characterized by the bottleneck with the maximum net flux [63]. In this way, it is possible to characterize the most probable pathways that characterize a complex system and to simplify the understanding of its dynamics.

4 Replica of previous results on a RhoGEF oncoprotein

4.1 Introduction

Ras homologue (Rho) are a family of proteins in the Ras superfamily of GTPase. These proteins are involved in many cellular functions, mainly in the dynamics of actin cytoskeleton and, consequently, in vesicles formation, membrane trafficking, and cell spreading [14]. The activity of small GTPases is regulated by GPTase-activating proteins and guanine-nucleotide exchange factors (GEF). The function of the latter is transiently stabilizing the nucleotide-free conformation of GTPases to allow the exchange between GDP and GTP inside the catalytic pocket [14], [15]. The most common structure of Rho GEF is characterized by a DH domain followed by a PH domain. The dynamics of such proteins have been analysed to investigate the relationship between their structure, its evolutionary-driven deformations, and their biological functions. The main functional specialization of this family of Rho GEF was discovered to be the autoinhibition by PH domain. Indeed, the collective motion of PH region was different between proteins in which the Rho GTPase binding surface is masked by the PH motif itself and the ones where this function is missing or carried out by other domains [52].

Rho guanine nucleotide exchange factor 12 (ARHGEF12), also known as leukaemia associated Rho GEF (LARG), is a Rho GEF whose in-frame fusion with MLL gene has been associated with acute myeloid leukaemia [64]. This protein is characterized by multiple domains and acts as an exchange factor for RhoA GTPase through its DH/PH region. LARG has been shown to mediate the activation of RhoA signalling by Gα-coupled receptors [65] and its enforced expression has been related to reduced cell proliferation and migration in breast and colorectal cancer [66]. Therefore, ARHGEF12 has been proposed as a potential tumour suppressor gene.

The dynamics and mechanical properties of LARG have been studied both in its free state and in presence of RhoA [34]. With the aim of finding a robust experimental setup to exploit in the analysis of Alsin DH/PH domain, the same systems were studied and our results were compared to the ones obtained from literature. At the same time, the main regions involved in the interaction of LARG with RhoA are investigated to carry out a successive comparative analysis with Alsin. In this way, it will be possible to characterize at a molecular level the different functions of ARHGEF12, a RhoGEF, and Alsin, a Rac1 effector.

4.2 Materials and Methods

4.2.1 Molecular Dynamics

The structures of ARHGEF12 alone (PDB: 1TXD) and bound to RhoA (PDB: 1X86) were retrieved from the Protein Data Bank [67]. For consistency with previous literature residue have been numberd according to UniProt entry Q9NZN5. Two MD simulations were performed, one for the protein without the GTPase (UnBnd system) and one for the protein bound to RhoA (Bnd system), using GROMACS 2020.4 [68]. AMBER ff99SB-ILDN force field was used to define the topology [69]. Both systems were configured in GROMACS in a cubic box with periodic boundary conditions setting a minimum distance of 1 nm between the protein and the box edge. Then, they were solvated in explicit TIP3P water [70] and, subsequently, an appropriate number of Na^+ and Cl^- were added to reach a physiological concentration of 0.15 M and to neutralize the charge. The energy minimization was performed through the steepest descend method for 2000 steps before equilibrating the systems. To this purpose, the following procedure was performed in both of them. An initial simulation of 500 ps in NVT ensemble and a following one of 500 ps in NPT ensemble were carried out restraining C-alpha carbons positions. The NVT simulation was performed with position restraints at a reference temperature of 300 K using the modified Berendsen thermostat [71] with $\tau = 0.1$ ps. The NPT simulation was carried out at 1.0 bar under position restraints using the Berendsen barostat with isotropic coupling and $\tau = 1.0$ ps. Finally, an MD simulation in NPT ensemble was produced for 260 ns. The equation of motion was integrated with the leap frog algorithm using a time step of 2 fs. Electrostatic interactions were treated with particle mesh Ewald method, , short-range cut-off at 1.2 nm and a switching of the potential starting at 1.0 nm. Van der Waals interaction were treated with a cut-off at 1.2 nm and a switching of the potential starting at 1.0 nm. The Visual Molecular Dynamics (VMD) engine was used for the visual inspection of systems and trajectories [72].

4.2.2 Analysis

The stability of each system was evaluated computing the root-mean-square deviation (RMSD) from the iniziato configuration of C-alphas atomic positions throughout the trajectory. Since the dynamics of the protein has been previously described as characterized by a collective motion of PH domain, the RMSD was also evaluated on the C-alphas of the sole DH region (residues 766-996). From the visual inspection of RMSD plots (Figure S1), last 240 ns of each trajectory was used in the following analysis. The flexibility of the protein was evaluated

computing the root-mean-square fluctuation (RMSF) during the last 240 ns and fitting the structures on the C-alphas of the DH domain.

To identify the residues of ARHGEF12 involved in the interaction with RhoA, the probability to be in contact with RhoA was computed in the Bnd system for each amino acid. The contact probability was computed sampling the MD trajectory every 250 ps with the following procedure [73]. For each sample snapshot, the distances between the atoms of one ARHGEF12 residue and the atoms of RhoA were computed: the residue was in contact if at least one of the residue-residue distances was lower than a threshold of 0.3 nm. The number of snapshots in which a residue was in contact divided by the total number of snapshots was the contact probability for that residue.

Previously, the force constant per residue profile was investigated in the same protein to infer its mechanical properties at the single residue level. This is a measure of the fluctuations of the mean distance of each residue from the rest of the structure and its higher values has been associated with protein functional sites [74], [75]. The calculation of force constants was implemented according to the formula:

$$k_i = \frac{3k_B T}{\langle (d_i - \langle d_i \rangle)^2 \rangle} \tag{38}$$

where d_i is the mean distance of the i-th residue from the rest of the structure, $k_B$is the Boltzmann's constant, T is the temperature of the system, and the operator $\langle\ \rangle$ stands for the average over the simulation. The distances were defined between the C-alphas of the amino acids and computed on representative snapshots extracted every 50 ps. The force constants were computed independently for the DH domain (residues 766-996) and the following region (residues 997-1126), comprising PH domain and the linker region.

Principal component analysis (PCA) was performed in both systems to analyse whether the presence of RhoA alters the essential dynamics of the RhoGEF. The covariance matrix was built on the C-alphas using the ones of DH domain for least square fit as done before [34].

GROMACS built-in tools have been used to compute RMSD, RMSF, and to perform PCA, while the calculation of contact probabilities and force constants was implemented using python libraries and custom scripts [76], [77].

4.2.3 Plots and Figures

Three-dimensional representations of the proteins were rendered in VMD. The principal directions were depicted through porcupine plots obtained from a custom made VMD script. In the porcupine plots, each C-alpha is associated to a segment oriented along the principal direction. The length of such segments is proportional to the amplitude of fluctuations along the represented direction. Data plots for RMSD, RMSF, force constants, and contact probability were generated using matplotlib library [78].

4.3 Results

Results will be compared to previous literature on a similar system [34].

4.3.1 RhoA interaction and mechanical properties

The DH domain of LARG, as the one of other RhoGEF proteins, is characterized by six main α-helices (α1-α6) organized in an oblong bundle. Within this structure, the three conserved regions are located respectively on α1, α2, and α5; while CR1 and CR3 form the GTPase binding surface, CR2 is exposed in the opposite side of the domain. The PH region is made of seven β-strands organized in an antiparallel way and followed by a C-terminal helix. The region connecting these two domains is characterized by multiple random coil sections (Figure 16).

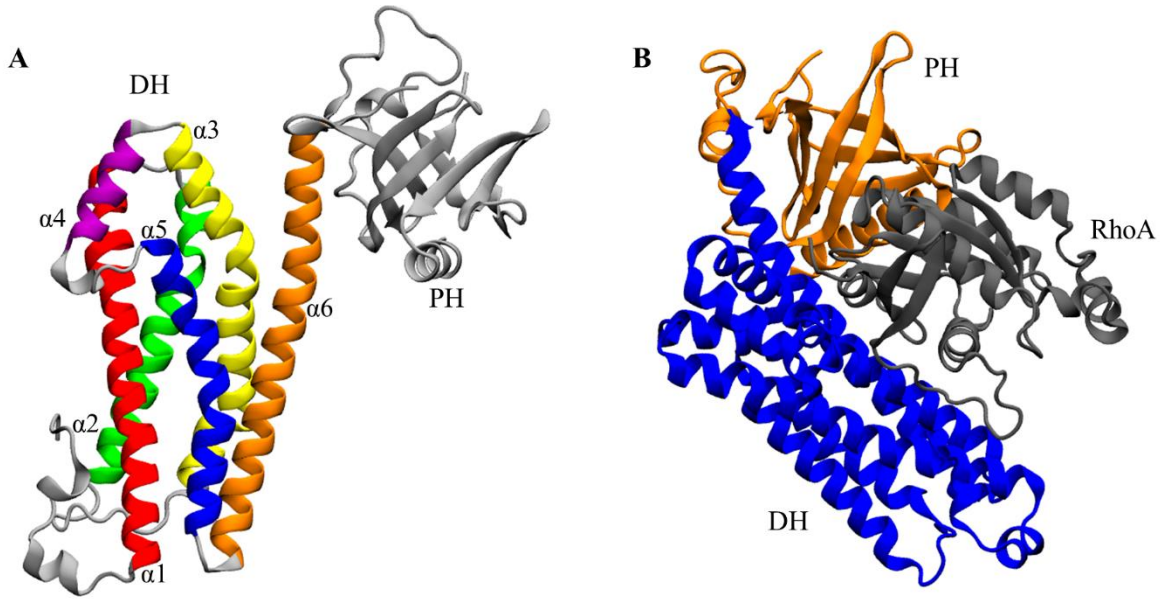

Figure 16. Crystal structures of LARG [67]. *(A) UnBnd state (PDB: 1TXD). Helices α1, α2, α3, α4, α5, and α6 of DH domain are coloured in red, green, yellow, purple, blue, and orange, respectively. (B) Bnd state (PDB: 1X86). DH domain, PH domain, and RhoA are coloured in blue, orange, and grey, respectively.*

Four main regions were involved in the interaction between LARG and the GTPase (Figure 17). Helices $\alpha 1$ and $\alpha 5$, corresponding to CR1 and CR3, were characterized by a high probability to be in contact with RhoA, supporting previous findings that those regions are crucial in the GEF activity of this family of proteins. Moreover, the central part of the helix $\alpha 6$ interacted with RhoA through almost all the dynamics suggesting its role in the regulation of the GTPase catalytic activity. Finally, there were some residues in helices $\alpha 3$ and $\alpha 4$ and in the non-structured region between $\alpha 4$ and $\alpha 5$ with a high contact probability, therefore these regions may have an auxiliary function in RhoA binding. Notably, PH domain did not show amino acids with a significant probability to be involved in the interaction despite being close to RhoA in the starting configuration.

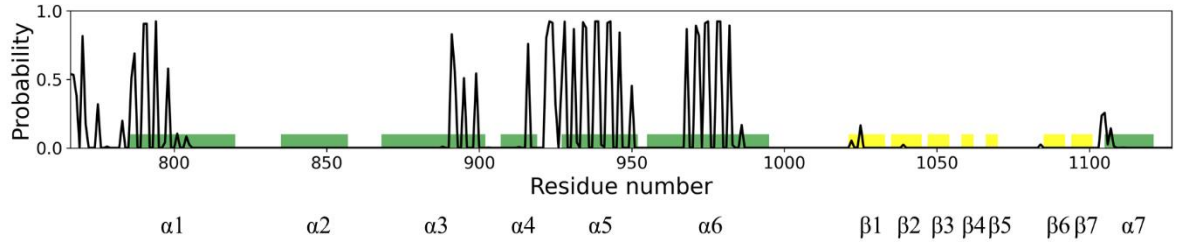

Figure 17. Contact probability for each residue of LARG. The secondary structure of DH and PH domains is highlighted to emphasize α-helices (green) and β-strands (yellow).

Previously, the mechanical profile of this protein was investigated. Independently of the functional state, the higher values of the force constants were located within the structured regions. Moreover, the peaks of the profiles corresponded to highly conserved residues within DH/PH domains. Finally, the presence of RhoA increased on average the mechanical rigidity of the protein. The same analysis was performed to investigate whether the different experimental setup could have had an influence on the dynamics of LARG. The force contstants obtained from our MD simulations were comparable to the ones from literature, both in the unbound (Figure 18) and in the bound (Figure 19) state. Moreover, in line with the previous findings, the mechanical profile of UnBnd and Bnd systems was similar.

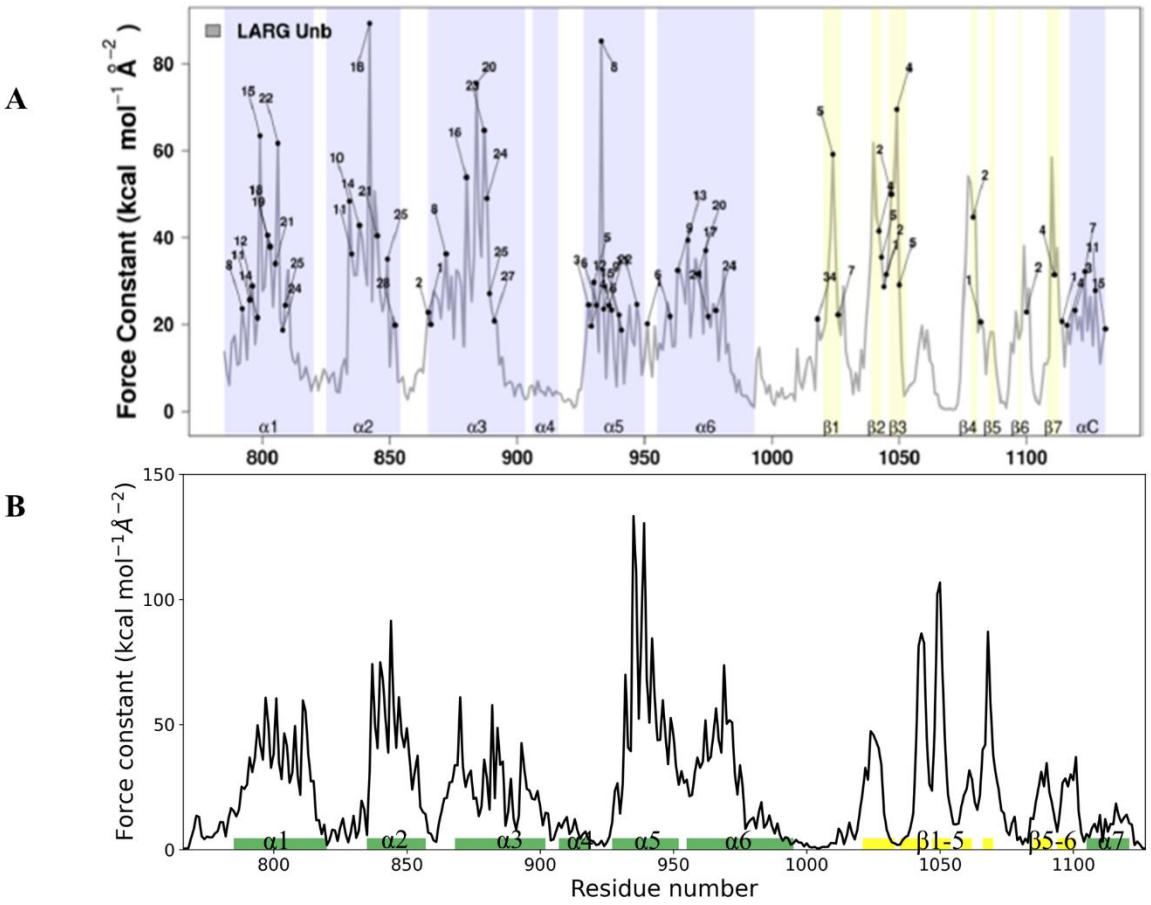

Figure 18. Comparison between the mechanical profile of UnBnd LARG from previous literature (A) and from our MD simulations (B). The secondary structure of DH and PH domain is highlighted to emphasize α-helices and β-strands. (A) Helices and strands are coloured in blue and yellow, respectively. (B) Helices and strands are coloured in green and yellow, respectively.

Independently of RhoA presence, the residues with the highest force constants were located in the helix $\alpha 5$. Here, the interaction with RhoA induced a remarkable increase of rigidity in the Bnd system, especially evident in the N-terminus of the region. The same effect could be observed within helix $\alpha 1$, except for the C-terminus where the values were comparable in the two functional states of LARG. The presence of RhoA did not cause significant differences within the last part of helix $\alpha 3$, while induced only a slight reduction in the fluctuations of helix $\alpha 4$. Moreover, helix $\alpha 6$ was characterized by low force constant values with no difference between the Bnd and UnBnd states despite interacting with RhoA. Therefore, the greatest change in the mechanical profile of the protein when it bound RhoA was represented by an increase of force constants within CR1 and CR3, which are crucial for the GEF activity of ARHGEF12. Despite showing some different values, on average the profile was similar to the one previously obtained and the main alterations due to the interaction were located in the same structured regions.

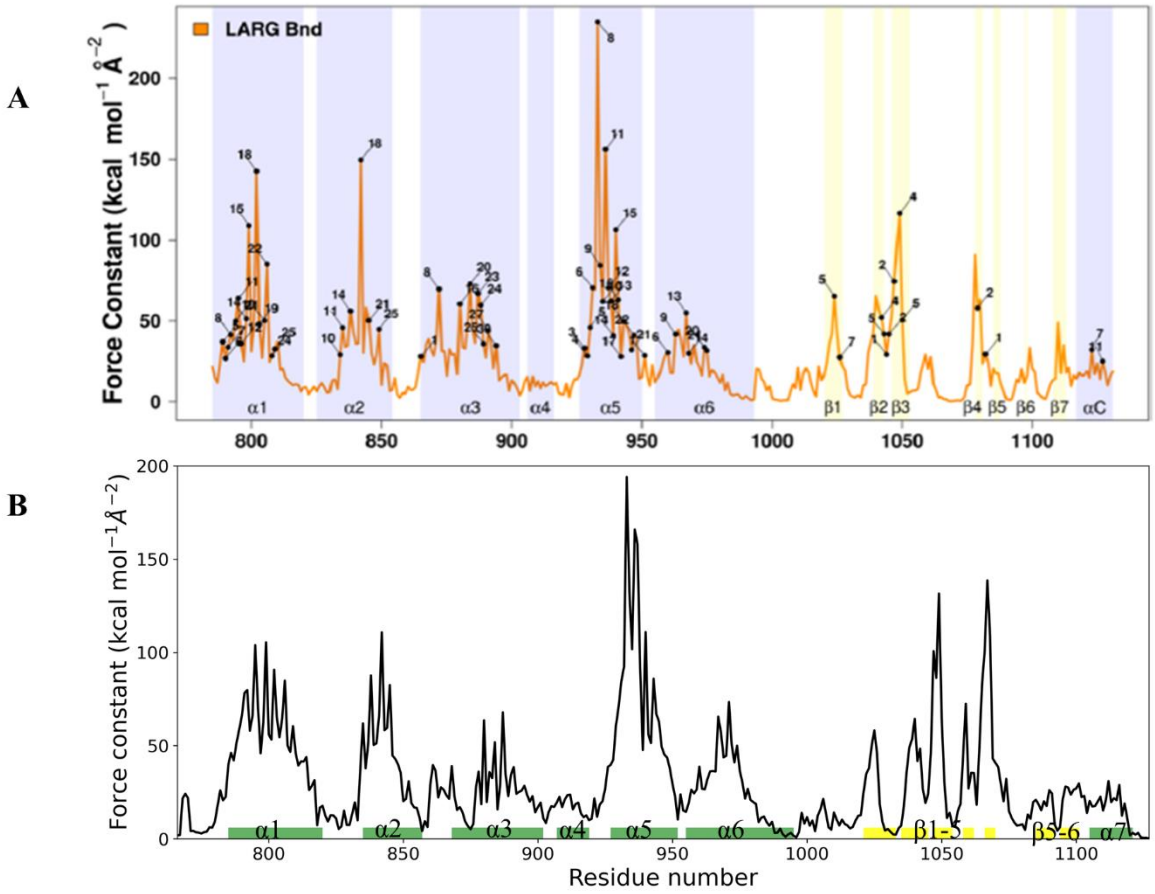

Figure 19. Comparison between the mechanical profile of Bnd LARG from previous literature (A) and from our MD simulations (B). The secondary structure of DH and PH domain is highlighted to emphasize α-helices and β-strands. (A) Helices and strands are coloured in blue and yellow, respectively. (B) Helices and strands are coloured in green and yellow, respectively.

The flexibility of LARG in the two states was initially evaluated through the RMSF on C-alphas. To compute this measure, protein structures were fitted on the C-alphas of the DH domain as was done previously in literature [34], [52]. The analysis on a similar system had revealed that the most flexible part of the protein is located mainly in the region following DH domain, with peaks corresponding to the loops connecting the β-strands of PH domain and the one linking DH and PH domains. Moreover, the fluctuations of the protein in two functional states were comparable with only a slightly increased mobility of free ARHGEF12. According to the results on these systems, the RMSF profile revealed that the two functional states have comparable flexibilities, but the Bnd system showed slightly greater fluctuations. In both cases, the essential motion resides in PH domain and the linker region, with the involvement of the last residues of helix α6. Moreover, it was possible to observe a reduction in the motion of helix α4 and the subsequent loop in the Bnd system, in agreement with the previous findings of its

interaction with RhoA. Despite small differences, the flexibility profiles of these systems were similar to the ones previously obtained, meaning that the different experimental setup had not influenced LARG dynamics, independent of the functional state (Figure 20).

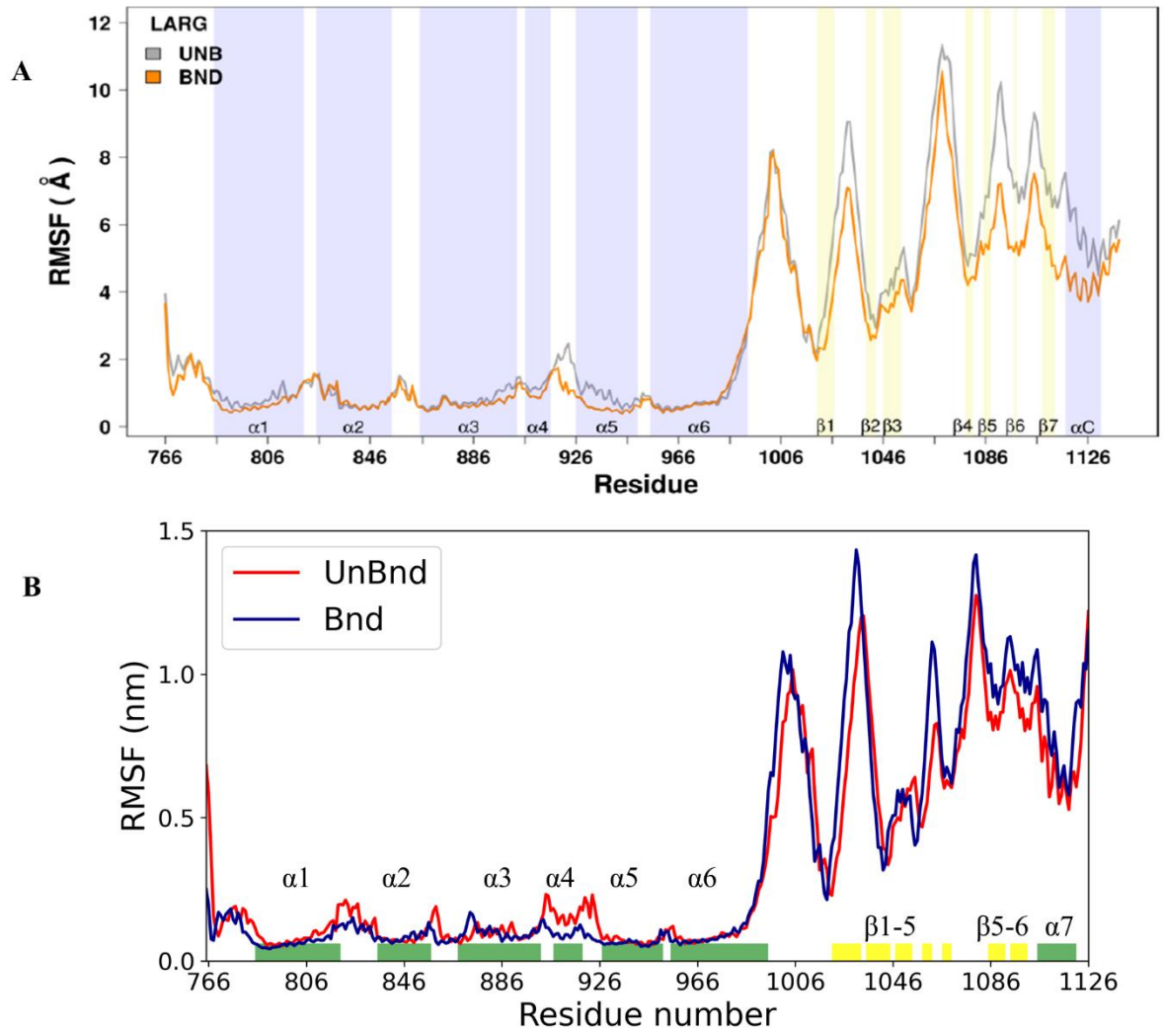

Figure 20. Comparison between the flexibility profile of free and bound LARG from literature and from MD simulations. The secondary structure of DH and PH domains is highlighted to emphasise α-helices and β-strands. (A) RMSF of bound and unbound LARG from literature, where helices and strands are coloured in blue and yellow, respectively. (B) RMSF of bound and unbound LARG from MD simulations, where helices and strands are coloured in green and yellow, respectively.

4.3.2 Analysis of the dynamics

To understand the essential dynamics that characterizes RhoGEFs, PCA has been applied to MD trajectories of several RhoGEF oncoproteins characterized by the DH/PH structure. These analysis had shown that, in all functional states, more than 80% of the total variance was represented by the first two principal components (Figure S2), which describe a collective

motion of PH domain and the terminal region of helix $\alpha 6$. Moreover, there was a significant overlap between these two directions between the bound and unbound states of the same protein [52]. Here, PCA has been used to investigate whether such collective motions were detectable also in these systems.

In agreement with the previous findings, the first two principal components (PC1 and PC2) explained almost the totality of the variance. In both functional states, PC1 and PC2 represented two different rotational motions of PH domain, together with the linker region and the last residues of DH domain (Figure 21). As previously observed, there was a remarkable overlap between the principal components of the two functional states. However, these results showed that, while in the UnBnd system PH domain moved closer to RhoA binding surface, in the Bnd system the dynamics followed the same directions but in the opposite way.

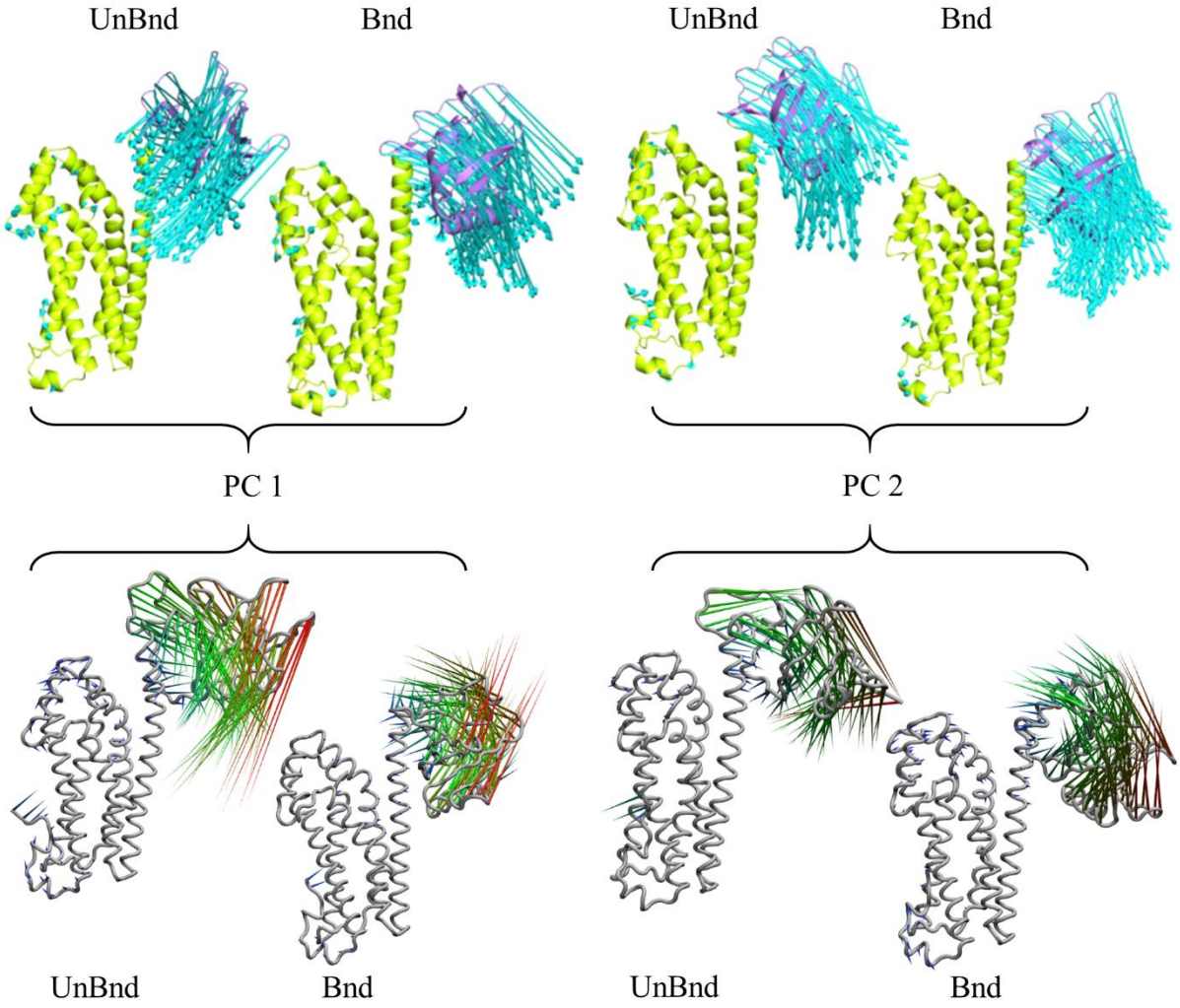

Figure 21. Comparison between the principal components of UnBnd and Bnd LARG from literature (top) and MD simulations (bottom).

4.4 Discussion

In this chapter, the dynamics of a known RhoGEF protein has been analysed in its free and RhoA-bound form and the results have been compared to previous findings on similar systems [34]. The close affinity of the obtained results with earlier literature provided a strong proof that the employed experimental setup was able to model the dynamics of proteins characterized by the DH/PH motif. Indeed, the mechanical profile of LARG has been reproduced and force constants values were similar to the ones previously obtained (Figure 18, Figure 19). The presence of RhoA increased on average the mechanical rigidity of the protein, especially in the regions involved in the interaction and in the loop connecting the DH and PH domains. Moreover, the C-terminal part of helix $\alpha 6$ was characterized by low rigidity independent of the functional state of LARG. In agreement with previous literature, the essential dynamics of this protein was a collective motion of PH domain, the region linking DH and PH regions, and the last part of helix $\alpha 6$ (Figure 20). The amount of fluctuation was similar in the bound and unbound states, meaning that the flexibility of the protein was not altered by the presence of its ligand partner. While previously slightly greater flexibility has been observed for free LARG, in the studied systems the PH domain was lightly more mobile in the bound form. Finally, it was possible to observe increased fluctuations in the region around helix $\alpha 4$ in absence of RhoA. Moreover, the PCA has showed that the essential dynamics could be described through two different roto-transational motions of PH domain (Figure 21). In fact, around 80% of the total variance could be expressed by the two first principal components (Figure S2). Notably, these directions were not altered by the presence of RhoA. However, the versus of motion in the Bnd system was opposite to the one in UnBnd system, while this difference was not observed before.

The information about regions involved in the interaction with RhoA given by the crystallographic structure is only partial. Indeed, the arrangement of proteins in a crystal may differ from the one in solution [79] and the dynamic information of the interaction is lost. Therefore, the probability for each residue of ARHGEF12 to be in contact with RhoA was investigated throughout the MD simulation (Figure 17). Despite being close in the initial configuration, the probability of having a contact between residues in PH domain and RhoA were very limited. The residues with highest probability to interact with the GTPase were located in CR1 and CR3, in agreement with their role in the GEF activity of this family of proteins. Besides, helices $\alpha 3$ and $\alpha 4$ might have had a role in stabilizing the interaction of RhoA with ARHGEF12 due to the presence of residues with high contact probability.

To conclude, the differences between the obtained results and the previous ones are contained within the range of variability of the employed methods. Moreover, the evidences were reproduced faithfully carrying out only one MD simulation instead of three, as done previously. Hence, the close similarity represented a validation of the employed experimental setup and the proof that it could be exploited for further analysis on Alsin DH/PH domain in order to strengthen the obtained results.

5 Analysis of Alsin dynamics

5.1 Introduction

IAHSP is a rare neurodegenerative disorder characterized by the onset of spasticity to the lower limbs within the second year of life and the progression towards tetraparesis [3]. The cause of this disease has been identified in the mutation of ALS2 gene, which encodes for Alsin protein [3], [4]. Its ability to interact with two GTPases, Rac1 and Rab5, is at the basis of its crucial role in vesicular trafficking, especially in neurons. In particular, Rac1 binding with DH/PH domain triggers the tetramerization and relocalization from the cytoplasm to the membrane, where Alsin acts as a Rab5 GEF [7]. Several studies have been carried out to investigate the physiological functions of this protein and how its mutation or loss leads to different forms of HSP [1], [2], [4], [12], [22], [24], [27], [29]. At the same time, the knowledge of the molecular mechanisms underlying Alsin biological functions and the nanoscale effect of mutations is crucial to design potential therapeutic strategies. However, an experimental structure of this protein has not been developed yet. To date, RLD is the only domain that has been modelled [4], [17], while none of Alsin regions has been studied exploiting MD tools.

The interaction between Rac1 and DH/PH domain is the first event of the pathways leading to the formation of early endosomes through Rab5 activation [7]. Indeed, Rac1 triggers a conformational transition of Alsin from a closed state, where RLD and the C-terminal region interact with each other, to an open state, in which it is able to form a tetramer [4]. Interestingly, the DH/PH motif is characteristic of a family of Rho GEF, but Alsin was demonstrated to be an effector rather than GEF [7]. Previously, the dynamics of homologous domains from other proteins has been investigated, showing that it consists essentially of a collective motion of PH domain and the last residues of DH domain [34], [52]. Given the fundamental role of this region in Alsin biological functions and its different role from the one of similar motifs, the aim of this work is to exploit homology modelling tools to build an atomistic model of Alsin DH/PH domain and characterize its dynamics, both alone and in presence of Rac1. To strengthen the results, the employed experimental setup has been tailored and validated replicating previous findings on LARG, a known Rho GEF oncoprotein [64], [66]. Moreover, a crystallographic structure showing this protein bound to its ligand partner, RhoA, has been used to model Alsin interaction with Rac1 [67]. First of all, Alsin dynamic has been compared to the one of LARG to understand the molecular basis of their different biological functions. Then, the conformations of Alsin DH/PH region were analysed both in presence and in absence of Rac1

to characterize the effect of such interaction at the nanoscale level. Finally, the dynamic of free Alsin was described through a Markov State Model to study the main states in which it could be found and discover the kinetics relationships between them. The results will provide an overall description of Alsin DH/PH domain possible conformations, both bound with Rac1 and alone. Furthermore, the molecular mechanism underlying the signal transduction between Rac1 and Alsin will be reported.

5.2 Materials and Methods

5.2.1 Homology Modelling

The three-dimensional structure of Alsin DH/PH domain was modelled starting from its amino acid sequence since no experimental structure is available. The amino acid sequence of human Alsin DH/PH domain (residues 686-1010) was downloaded from NCBI database. Then, the homology model was built giving it as input to the I-Tasser suite [57], [80], [81]. Among the output models, the one with the highest C-score was retained. The C-score is a parameter between -5 and 2 computed by I-Tasser considering the results of the different steps followed to develop the model itself. It is used to establish the level of confidence of the obtained structure, where higher values correspond to higher quality. From this score, I-Tasser also estimates the TM-score, which measures the structural similarity between two proteins, and the RMSD between the predicted model and the native structure. The secondary structure of the homology model was analysed through the STRIDE software package. The quality of the model was also evaluated in MOE [82] through the visualization of the Ramachandran plot, which represents the distribution of phi and psi angles pairs and the allowed regions. The percentage of residues lying in not allowed regions were compared to those of the templates used by I-Tasser during the construction of the model. Finally, for each template, the identity and similarity scores relative to Alsin were computed using MOE with the following procedure: the crystal structures of the templates were retrieved from Protein Data Bank, their sequences were aligned with the one of Alsin DH/PH domain, the amino acids outside the region covered by Alsin residues were deleted, and then the scores were computed dividing by the length of Alsin sequence. BLOSUM-62 score matrix was used to perform the alignment and compute the similarity scores.

5.2.2 Molecular Dynamics

Two systems were analysed, free Alsin DH/PH domain ($Alsin^{UnBnd}$) and Alsin DH/PH domain bound to Rac1 ($Alsin^{Bnd}$). To obtain the initial configuration of $Alsin^{UnBnd}$, the protonation state

of the homology model was adjusted at a physiological pH of 7.4 using MOE. The initial configuration of the Alsin^{Bnd} was obtained superimposing the homology model and Rac1 (PDB: 3TH5 [83]) to LARG and RhoA (PDB: 1X86 [67]), respectively. The nucleotide and magnesium ion were removed from Rac1 as they were not present in previously analysed systems [34]. Then, the protonation state was adjusted according to a physiological pH of 7.4 and, to avoid steric clashes due to superimposition, the energy minimized using MOE. Two MD simulations were performed for each system using GROMACS 2020.4 [68] and AMBER ff99SB-ILDN force field to define the topology [69]. The following procedure has been employed for both systems. Protein was inserted in a cubic box with periodic boundary conditions defined setting a minimum distance of 1 nm between the protein and the box edge. Then, it was solvated in explicit TIP3P water [70] and, subsequently, an appropriate number of Na^+ and Cl^- were added to reach a physiological concentration of 0.15 M and to neutralize the charge. The energy minimization was performed through the steepest descend method for 2000 steps. Then, two replicas were obtained from each system as follows. An initial simulation of 500 ps in NVT ensemble and a following one of 500 ps in NPT ensemble were carried out, both of them under position restraints. The NVT simulation was performed at a reference temperature of 300 K using the modified Berendsen thermostat [71] with $\tau = 0.1$ ps. The NPT simulation was carried out at 1.0 bar using the Berendsen barostat with isotropic coupling and $\tau = 1.0$ ps. Finally, a MD simulation in NPT ensemble was produced for 500 ns. The equation of motion was integrated with the leap frog algorithm using a time step of 2 fs. Electrostatic interactions were treated with particle mesh Ewald method, short-range cut-off at 1.2 nm and a switching of the potential starting at 1.0 nm. Van der Waals interaction were treated with a cut-off at 1.2 nm and a switching of the potential starting at 1.0 nm.

To better explore the state space in the free form of Alsin, four additional MD simulations, each one 100 ns long, were performed with the same procedure described before. The initial configurations were extracted from the trajectories of Alsin^{UnBnd} replicas. From now on, the two 500 ns long trajectories will be called “long replicas”, while these four “short replicas”.

With the aim of studying the transition from the bound to the unbound states, the structure of the sole DH/PH domain was extracted at 260 ns of the first long replica of Alsin^{Bnd}. Then, it was used to produce a new trajectory of 260 ns with the same experimental setup described before. In the following analysis, the trajectories before and after Rac1 removal will be called Alsin^{Rac1} and Alsin^{noRac1}.

The Visual Molecular Dynamics (VMD) engine was used for the visual inspection of systems and trajectories [72].

5.2.3 Analysis

The stability of each system was evaluated through the root-mean-square deviation (RMSD) from the initial configuration of C-alphas atomic positions during the trajectory. Since previously it has been observed that the essential dynamics of DH/PH domains in other proteins is characterized by a collective motion of PH domain, the RMSD was computed also to the C-alphas of the sole DH domain (residues 686-895). From the visual inspection of RMSD plots (Figure S3), last 450 ns of each long trajectory was considered in the following analysis. As for the short trajectories, the last 90 ns were used in the analysis. The RMSF, force constants, and contact probability were computed as described in the previous chapter (see section 4.2.2) for each long replica, then the results were averaged. To compute the mechanical properties at residue level, DH domain (residues 686-895) and the following region (residues 896-1010), comprising PH domain and the non-structured linker, were considered independently.

The position of PH domain with respect to DH region in the bound and unbound states has been investigated to understand the effect of Rac1 in the conformations of Alsin DH/PH domain. To this purpose, two coordinates were defined using a DH-based reference system. Identifying as x and y axis the first and second principal directions of the DH domain, the z axis is the one perpendicular to plane xy. α_{xy} has been defined as the angle in plane xy between the straight line parallel to the x axis passing through DH domain (residue 686-864) centre of mass and the segment linking the latter to PH domain (residues 914-1010) centre of mass. Since the motion of last helix in DH region is involved in PH dynamics in other proteins, it was not considered when computing the centre of mass to avoid possible changes in its position due to PH fluctuations. Then, the coordinate d_z was defined as the distance along z axis between the centres of mass of DH and PH domains, such that if d_z is positive the latter is above the DH region centre of mass, and therefore closer to Rac1-binding surface. Figure 22 shows a graphical representation of the employed coordinates.

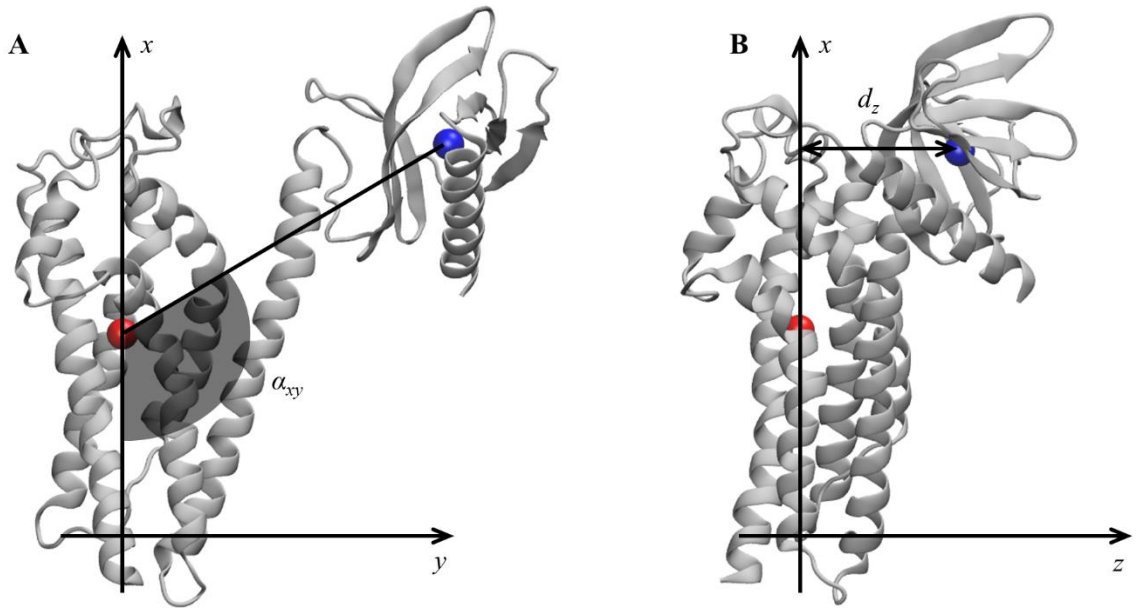

Figure 22. Coordinates used to describe the relative position between DH and PH domains. Centred of mass of DH and PH domains are represented as red and blue spheres, respectively. (A) α_{xy} is the angle in plane xy between x axis and the segment bridging the centres of mass. (B) d_z is the distance measured along z axis between the two centres of mass. In this figure, the distance is negative because PH centre of mass is under DH centre of mass, i.e. has greater z coordinate.

In the analysis of PH-DH relative position, the short replicas were considered in order to obtain a wider sampling of the state space. To compute the potential mean force (PMF) along these two directions, the probability $p(\alpha_{xy}, d_z)$ of the system to be in the point (α_{xy}, d_z) was obtained from the bidimensional histogram $H(\alpha_{xy}, d_z)$ of Alsin$^{\text{UnBnd}}$ and Alsin$^{\text{Bnd}}$ ensembles as:

$$p(\alpha_{xy}, d_z) = \frac{H(\alpha_{xy}, d_z)}{\sum_{\alpha_{xy}} \sum_{d_z} H(\alpha_{xy}, d_z)} \tag{39}$$

where bins of 1° and 0.1 nm were used to discretize the state space along α_{xy} and d_z directions, respectively. Boltzmann inversion was then performed to obtain the PMF along the two coordinates:

$$PMF(\alpha_{xy}, d_z) = -k_B T \ln p(\alpha_{xy}, d_z) \tag{40}$$

where k_B is the Boltzmann's constant and T is the temperature.

To evaluate the effect of the interaction between Rac1 and helix α6 (residues 865-895) on its straightness, the curvature of α6 axis on plane *xz* was analysed as follow [84]. Representative snapshots were extracted every 50 ps for both long and short replicas. For each snapshot, the *x*

and z coordinates of the alpha carbons were picked, then the centres of mass of successive groups of four C-alphas were considered as points of helix axis. Therefore, the i-th sample of the axis is obtained selecting from the i-th to the $(i+3)$-th alpha carbons and computing their centre of mass. The obtained points were interpolated, using the x coordinate as independent variable, with a second-degree polynomial function $c(x)$ which was then evaluated in 100 points to approximate the helix axis (Figure 23).

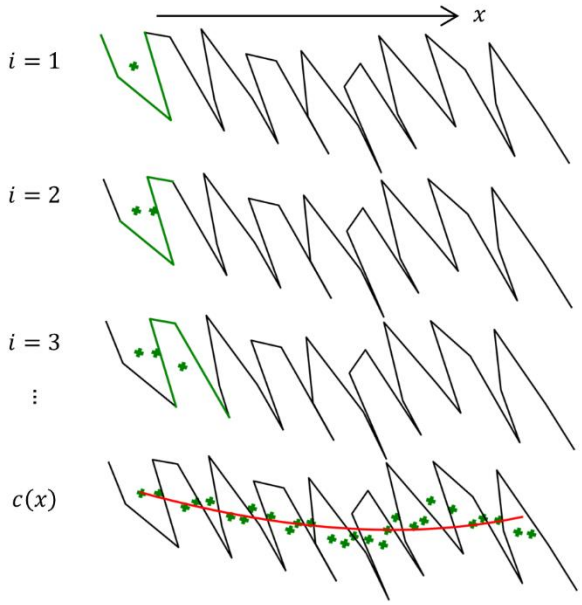

Figure 23. Approximation of the helix axis. Successive groups of four atoms are considered and the centre of mass is computed for each of them. Then, the axis is represented by a second-degree polynomial function $c(x)$ obtained through the interpolation of the obtained points.

Finally, the curvature $\kappa(x)$ and the integral of curvature I_κ were computed as:

$$\kappa(x) = \frac{|c''(x)|}{(1 + c'(x)^2)^{3/2}} \qquad I\kappa = \int \kappa(x)\, dx \tag{41}$$

The integral was numerically solved using the composite trapezoidal rule. Higher values of integral of curvature are related to higher deviations from the straightness of helix axis.

To analyse the effect of Rac1 interaction on helix $\alpha 3$ position and, therefore, on the region between it and helix $\alpha 5$ (α3-5), two quantities were computed extracting representative snapshots every 50 ps for both long and short replicas. The first one is the distance along z axis between DH domain centre of mass, without considering the helix $\alpha 6$ as in the previous analysis, and the region of helix $\alpha 3$ in contact with Rac1 (residues 788-793), such that positive values indicate $\alpha 3$ being over DH centre of mass. The second one is the distance between the centres of mass of α3-5 (residues 796-816) and PH domain. As for the comparison between

Alsin^{Rac1} and Alsin^{noRac1}, the last 100 ns of their trajectories were considered in the analysis of $\alpha 3$ position and $\alpha 3$-5 distance from PH region.

The conformational dynamics of Alsin^{UnBnd} was investigated through a Markov State Model (MSM) to discover the kinetic relationships between the main accessible states. The state space was described in terms of DH-PH relative position through the previous mentioned coordinates, which were computed every 10 ps for both long and short replicas [58]. To finely discretize the space state, data were divided into 1000 clusters using K-centres algorithm. Sliding window method and maximum-likelihood estimation were used to obtain the count matrix and the transition matrix, respectively. The optimal lag time was chosen analysing the largest implied timescale at lag times between 0.5 ns and 17.5 ns. Then, to better understand the obtained model, the microstates were divided into a smaller number of states using the Robust Perron Cluster Cluster analysis (PCCA+) algorithm [60]. The number of states was chosen observing the distribution of the slowest ten implied timescales at the chosen lag time. A new MSM was estimated using the clusterization obtained from PCCA+ and, then, was validated through a Chapman-Kolmogorov test performed as follows. If $[T(\tau)]^k$ is the k-fold application of the transition matrix of the model to validate, i.e. the matrix multiplied by itself k times, and $T(k\tau)$ the transition matrix at the greater lag time $k\tau$, the approximation (32) was tested estimating $T(k\tau)$ through a *Markov chain Monte Carlo sampling* of 100 transition matrices from 2 independent Markov chains. In particular, for a given lag time the mean of the obtained distribution was used as transition matrix and the standard deviation was taken as a measure of statistical uncertainty [59], [61], [62]. The test was performed for $k > 2$ and $k\tau < 50$ ns to avoid excessive under sampling of short trajectories. Finally, transition path theory was applied to the validated model to identify the most probable transition pathways characterizing free Alsin DH/PH domain. Once the initial and final states of interest were chosen, the net flux matrix was computed. Then, it was used to find the most probable pathway from the starting to the final state through the Dijkstra algorithm, which finds the path characterized by the bottleneck with the highest flux.

GROMACS built-in tools were used to compute RMSD and RMSF. Contact probabilities, force constants, the coordinates defined to describe the system, and PMF were obtained using python libraries and custom made scripts [76], [77]. MSMBuilder libraries were used to build the MSM and perform transition path theory [85].

5.2.4 Plots and figures

Three-dimensional representations of the proteins were rendered in VMD. Ramachandran plots were generated in MOE, while all other data plots using matplotlib library [78]. The network representation of the transition matrix was obtained through PyEMMA libraries [86].

5.3 Results

5.3.1 Homology model of Alsin DH/PH domain

The homology model of Alsin DH/PH domain was built by I-Tasser using 16 templates. The best model was characterized by a C-score of 0.66, an estimated TM-score of 0.80 ± 0.09, and an estimated RMSD of 5.0 ± 3.2 Å. On average, the sequence identity and similarity between the templates and Alsin were 12.5% and 28.2%. The quality of the structure was also investigated by observing its Ramachandran plot (Figure 24), computing the percentage of residues lying in not-allowed regions, and comparing this value with the average on the templates. The percentage of Ramachandran outliers in Alsin homology model and the average on the chosen templates were 1.5% and 0.8%, respectively. Therefore, from the analysis of the torsional angles, the quality of Alsin model is in line with the crystallographic structures of the templates. The PDB identification codes, identities, similarities, and percentages of Ramachandran outliers of the templates are listed in Table 2.

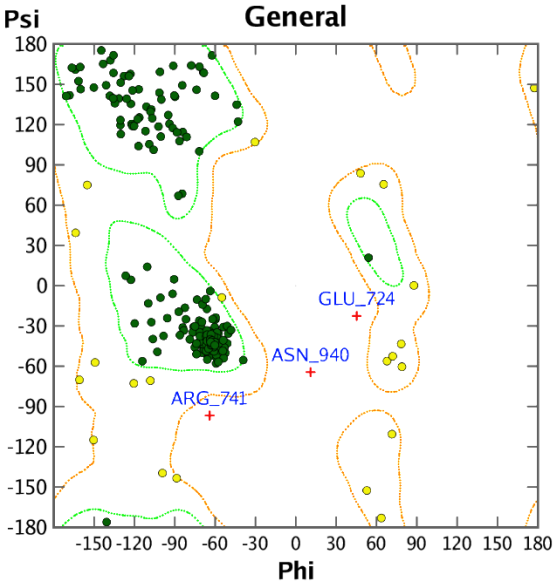

Figure 24. Ramachandran plot of Alsin DH/PH homology model. Residues in the core regions, residues in allowed regions, and outliers are coloured in green, yellow, and red, respectively.

Table 2. Information about the templates used to build Alsin homology model.

PDB	Identity (%)	Similarity (%)	Ramachandran outliers (%)
1FOE [39]	16.9	29.2	2.5
1KI1 [87]	12.9	29.2	3.3
1NTY [88]	10.5	26.5	0.0
1XCG [89]	14.2	28.3	1.5
1X86 [67]	13.5	29.8	0.3
2DFK [90]	10.2	27.7	0.6
2PZ1 [91]	11.4	30.5	2.4
2RGN [92]	10.5	25.8	0.3
2Z0Q [93]	13.2	28.9	0.0
3MPX [94]	10.5	22.5	0.0
3ODO [95]	14.2	28.9	0.9
4D0N [96]	12.0	31.1	0.0
4GZU [42]	11.4	24.6	0.0
4XH9 [97]	12.9	28.9	0.3
4YON [35]	14.8	32.3	0.0
6D8Z [98]	11.1	26.5	0.0

As in other proteins, the DH domain of Alsin is characterized by six main α-helices (α1-α6) organized in an oblong bundle. While helices α1 and α5 are exposed on the same side of the domain, probably forming the Rac1-binding region as in other proteins, helix α2 is exposed in the opposite surface, which is involved in dimerization of other DH domains and, most likely, Alsin itself. The third helix of DH region exposes its N-terminus and C-terminus at the dimerization and Rac1-binding surfaces, respectively. Finally, the helix α6 is located on one side of the domain and is connected by a random coil region to the PH domain. The latter is composed by six antiparallel β-strands followed by one α-helix and is organized in a globular structure (Figure 25). Therefore, the structure of Alsin DH/PH domain is characterized by the same motifs of other Rho GEF proteins.

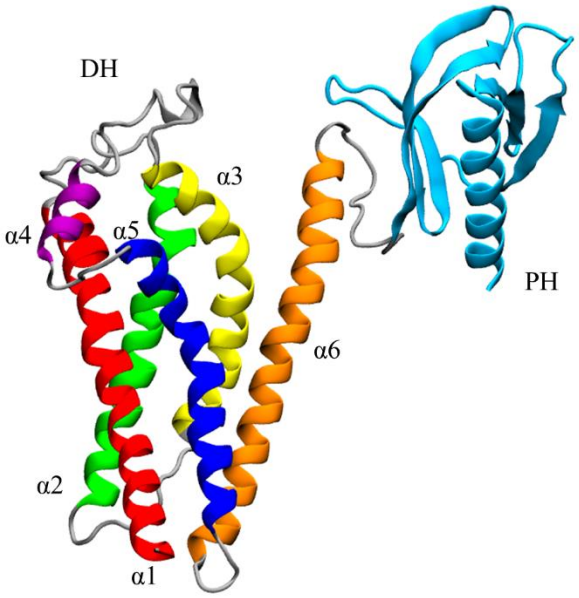

Figure 25. Homology model of Alsin DH/PH domain. The helices α1, α2, α3, α4, α5, and α6 of DH domain are coloured in red, green, yellow, purple, blue, and orange, respectively. PH domain is coloured in cyan.

To identify the conserved regions in Alsin DH/PH domain, its amino acid sequence was aligned in MOE with BLOSUM-62 score matrix with the ones of the 16 templates used to build the model. Then, the residues forming the conserved regions of TIAM1 [39] were used to locate them in Alsin (Figure S4). CR1 and CR3 were characterized by a higher number of conserved residues among the analysed proteins, while no amino acid was totally conserved in CR2. The secondary structure of the model, analysed through STRIDE software package, the residues composing helices and strands, and the conserved regions are showed in Figure 26.

α1
686 KNIMGYIASLHELATTERRFYSKLSDIKSQILRPLLSLENLGTTTTVQLLQEVASRFS
α2 α3
744 KLCYLIGQHGASLSSFLHGVKEARSLVILKHSSLFLDSYTEYCTSITNFLVMGGFQL
α4 α5
801 LAKPAIDFLNKNQELLQDLSEVNDENTQLMEILNTLFFLPIRRLHNYAKVLLKLAT
α6
857 CFEVASPEYQKLQDSSSCYECLALHLGRKRKEAEYTLGFWKTFPGKMTDSLRKPE
β1 β2 β3 β4
912 RRLLCESSNRALSLQHAGRFSVNWFILFNDALVHAQFSTHHVFPLATLWAEPLSEE
β5 β6 α7
968 AGGVNGLKITTPEEQFTLISSTPQEKTKWLRAISQAVDQALRG 1010

Figure 26. Definition of secondary structure for Alsin DH/PH domain. Helices are represented by green rectangles, coils by black lines, and strands by yellow arrows. The conserved regions CR1, CR2, and CR3 are highlighted in red, blue, and purple, respectively.

5.3.2 Rac1 interaction and mechanical properties

The initial configuration of Alsin^{Bnd} was modelled using the crystallographic structure of RhoA-bound LARG (Figure S5). The RMSD at the end of the superimposition were 1.98 Å between the DH/PH domains and 2.36 Å between the GTPases. The regions involved in the interaction of Alsin with Rac1 have been investigated computing the probability of each residue to be in contact with the GTPase (Figure 27). The amino acids with a contact probability greater than 0.9 were located in helices $\alpha 3$, $\alpha 5$, and $\alpha 6$ indicating that these are the main structures forming the Rac1-binding surface. Moreover, probabilities around 0.5 were found for three loops, the one immediately C-terminal to helix $\alpha 3$ (l1), the one between helices $\alpha 4$ and $\alpha 5$ (l2) and, in PH domain, the portion C-terminal to strand $\beta 3$ (l3). Notably, helix $\alpha 1$ was not involved in significant interactions with Rac1 unlike LARG despite the starting configuration of Alsin^{Bnd} system was obtained from the crystallographic structure of RhoA-bound LARG (see section 4.3.1). Thus, between the conserved regions that are responsible for the interaction with Rho GTPase, only Alsin CR3 interacted with Rac1.

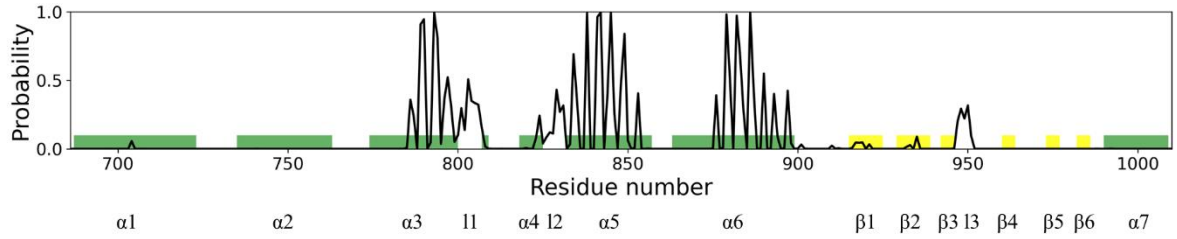

Figure 27. Alsin-Rac1 contact probability. The secondary structure of Alsin is highlighted to show helices (green) and strands (yellow).

The mechanical properties at the single residue level of Alsin^{Bnd} and Alsin^{UnBnd} were inferred computing the force constants and then were compared to analyse the effect of Rac1 interaction on Alsin. Consistently with similar domains, the highest values were located in the structured regions, the presence of Rac1 increases on average the rigidity of the domain, and the greatest force constants were obtained in the first part of helix $\alpha 5$, independently of the state (Figure 28). As for LARG (see section 4.3.1), the most evident effect of Rac1 interaction was the increased mechanical rigidity within the first half of helix $\alpha 5$. On the other hand, it was possible to observe only little changes in the mechanical profile of the first helix. Despite producing a

contained increase in the force constants within its first half, the presence of Rac1 slightly increased the fluctuations of following residues. Moreover, the second half of helix $\alpha 6$ and the region around helix $\alpha 4$ were characterized by higher force constants in presence of Rac1. Limited differences could be observed within PH domain, where in absence of Rac1 the rigidity of strand $\beta 4$ is lower while that of strand $\beta 6$ is higher. Finally, in the bound state the mechanical properties within helix $\alpha 2$ and the first part of $\alpha 3$ were increased even though these regions did not interact with Rac1. Therefore, in presence of Rac1 Alsin seems to increase the mechanical rigidity of residues exposed at the putative surface of dimerization.

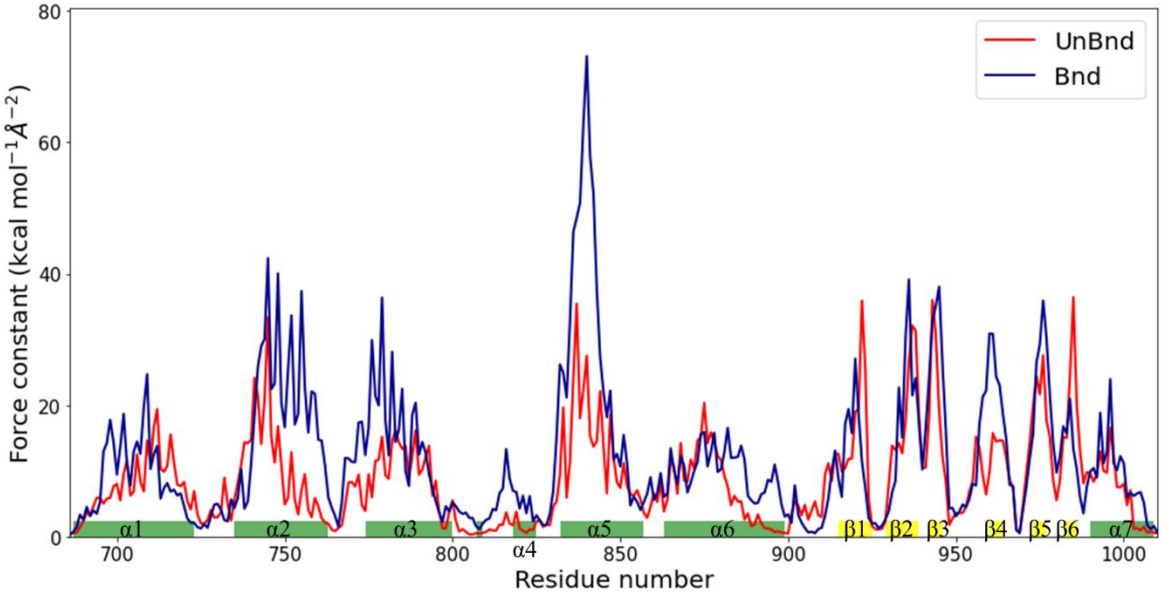

Figure 28. Mechanical profile of Alsin DH/PH domain in the UnBnd and Bnd states. Residues forming helices and strands are coloured in green and yellow, respectively.

The regions that characterize Alsin dynamics have been investigated through RMSF computed on C-alphas (Figure 29). As for LARG, the greatest fluctuations were located within PH domain both in presence and in absence of its binding partner (see section 4.3.1). However, in presence of Rac1 the flexibility of helix $\alpha 6$ was reduced indicating that the last residues of DH domain were less involved in the collective motion of PH region. On the other hand, the coiled coil linker between the two domains contributed more to the dynamics in Alsin^{Bnd} than in Alsin^{UnBnd}. Moreover, unlike LARG, it is possible to identify a region characterized by higher fluctuations between helices $\alpha 3$ and $\alpha 5$ (α3-5), with two peaks located before and after $\alpha 4$. Here, the flexibility is reduced in presence of Rac1. Finally, in the bound state the fluctuations within loop l3 were reduced with respect to the unbound state.

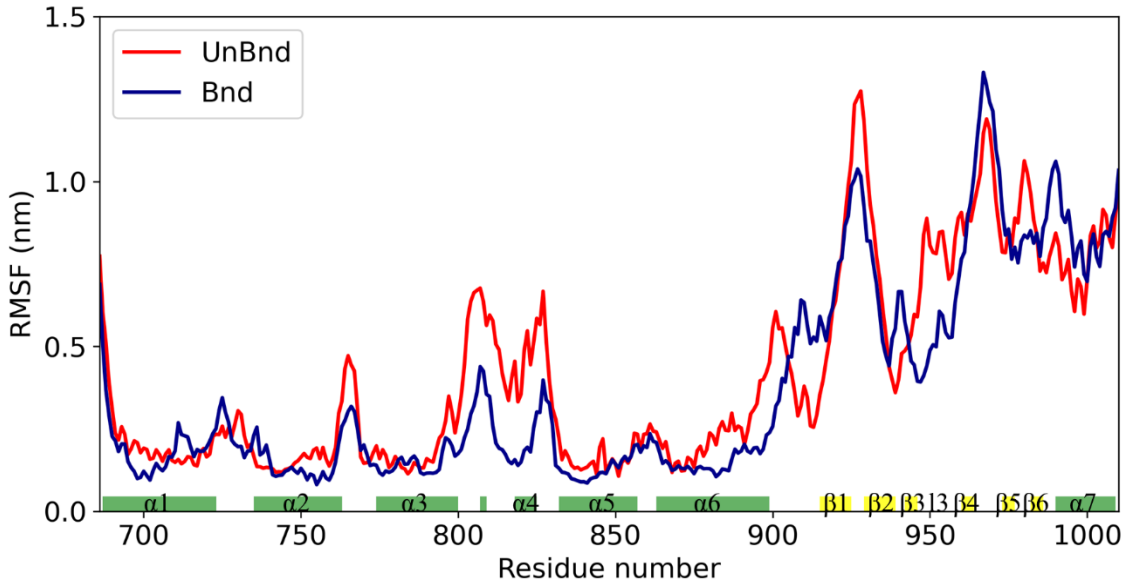

Figure 29. RMSF of free and Rac1-bound Alsin. Helices and strands are represented in green and yellow, respectively.

5.3.3 Effect of Rac1 interaction on PH dynamics

Since the essential dynamics of Alsin is a collective motion of PH domain with respect to DH domain, further investigations have been made to characterize their relative position in the bound and unbound states. To this purpose, the PMF along the two chosen coordinates, α_{xy} and d_z, was analysed (Figure 30). It was possible to observe that only in absence of Rac1 angles lower than 120° or distances lower than -2 nm were accessible. On the other hand, positive distances were obtained almost only in Alsin^{Bnd} with the only exception of the region $(\alpha_{xy} \approx 125°; d_z \approx 0)$, in which the free energy profile of the two states is similar. The overall ability to explore different DH-PH relative position is higher in Alsin^{UnBnd}, where the minima are well connected. On the other hand, in presence of Rac1 the minima are narrower and two low connected regions, characterized by positive and negative d_z, are explored. According to the definition of the two coordinates, the PMF showed that, in absence of Rac1, more closed conformations were explored by the domain, i.e. the positions assumed by PH domain tended to move Alsin C-terminus closer to the N-terminus. On the other side, the presence of Rac1 seemed to stabilize a more linear and open conformation of the domain.

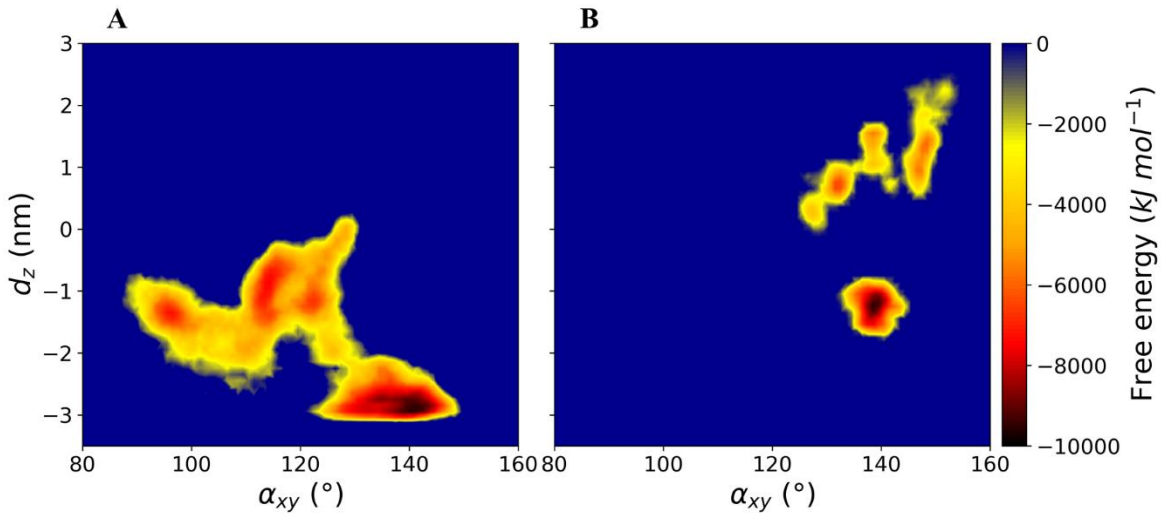

Figure 30. PMF along α_{xy} and d_z for (A) Alsin^{UnBnd} and (B) Alsin^{Bnd}.

The effect of the interaction between Rac1 and helix $\alpha 6$ has been investigated through the analysis of the helix axis curvature. The average curvature integral was significantly higher in Alsin^{UnBnd} (1.06 ± 0.34) than in Alsin^{Bnd} (0.22 ± 0.17) meaning that the presence of Rac1 stabilized a straighter conformation of $\alpha 6$ (Figure 31 A). Moreover, the curvature integral in the UnBnd state tended to be higher in those conformations characterized by lower values of α_{xy} or d_z, i.e. within regions in which the domain is in a closed state (Figure 31 B). At the same time, low values of curvature integral for Alsin^{UnBnd} were obtained in two regions, the one partially overlapping with Alsin^{Bnd} and the one characterized by $\alpha_{xy} \in [115°, 120°]$ and $d_z \in [-1, 0]$. Notably, the deepest minimum of the UnBnd state, in which the domain is closed, was characterized by some of the lowest values of curvature integral, while the one of the Bnd state by values approximatively nil. Therefore, through its action on the last helix of DH domain (Figure 31 C, D), Rac1 was able to alter the dynamics of PH domain.

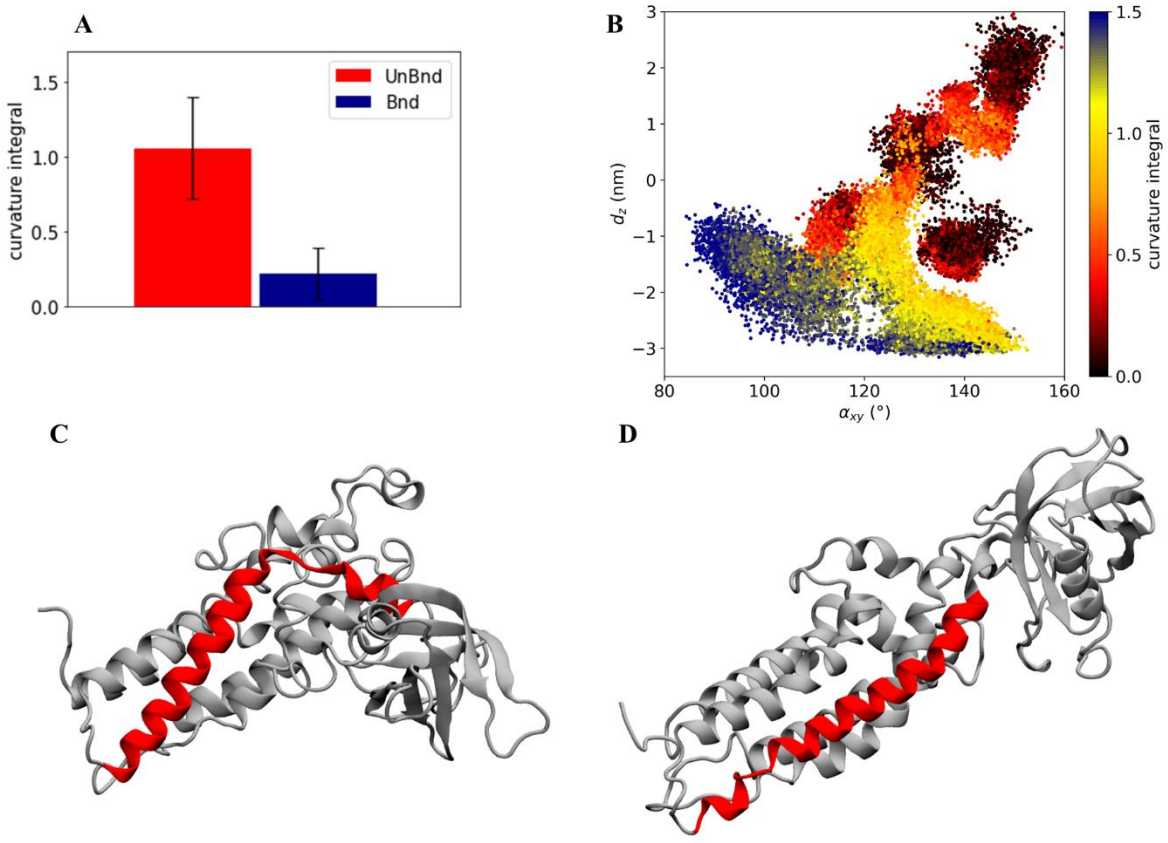

Figure 31. Analysis of helix α6 curvature. (A) Bar diagram showing the average curvature in the two states, where the error bars represent the standard deviation of the distribution. (B) Representation of the curvature integral value depending on the position of the protein in the α_{xy}-d_z plane. Each point is a snapshots of the trajectories and is coloured according to the level of curvature of helix α6 (C) Representative snapshot of Alsin^{UnBnd} where α6 is highlighted in red. (D) Representative snapshot of Alsin^{Bnd} where α6 is highlighted in red.

Finally, the effect of Rac1 in the position of helix α3 has been described in terms of distance along the *z* axis between its residues that were in contact with Rac1 and DH domain centre of mass (Figure 32 A). It was possible to observe greater fluctuations of this measure in Alsin^{UnBnd} than Alsin^{Bnd} throughout the dynamics. Moreover, the average distance was negative in the UnBnd state (-0.05 ± 0.16 nm), while positive in the Bnd state (0.19 ± 0.06 nm). The result of such Rac1-induced displacement has been described through the distance between α3-5 and PH domain (Figure 32 B). It was possible to observe a remarkable difference between the two states, with lower values in presence of Rac1 (2.74 ± 0.20 nm) than in case of free Alsin (3.90 ± 0.54 nm). Thus, the interaction with Rac1 stabilized the position of helix α3 above DH domain centre of mass and, as a result, the region α3-5 moved closer to PH domain (Figure 32 C, D).

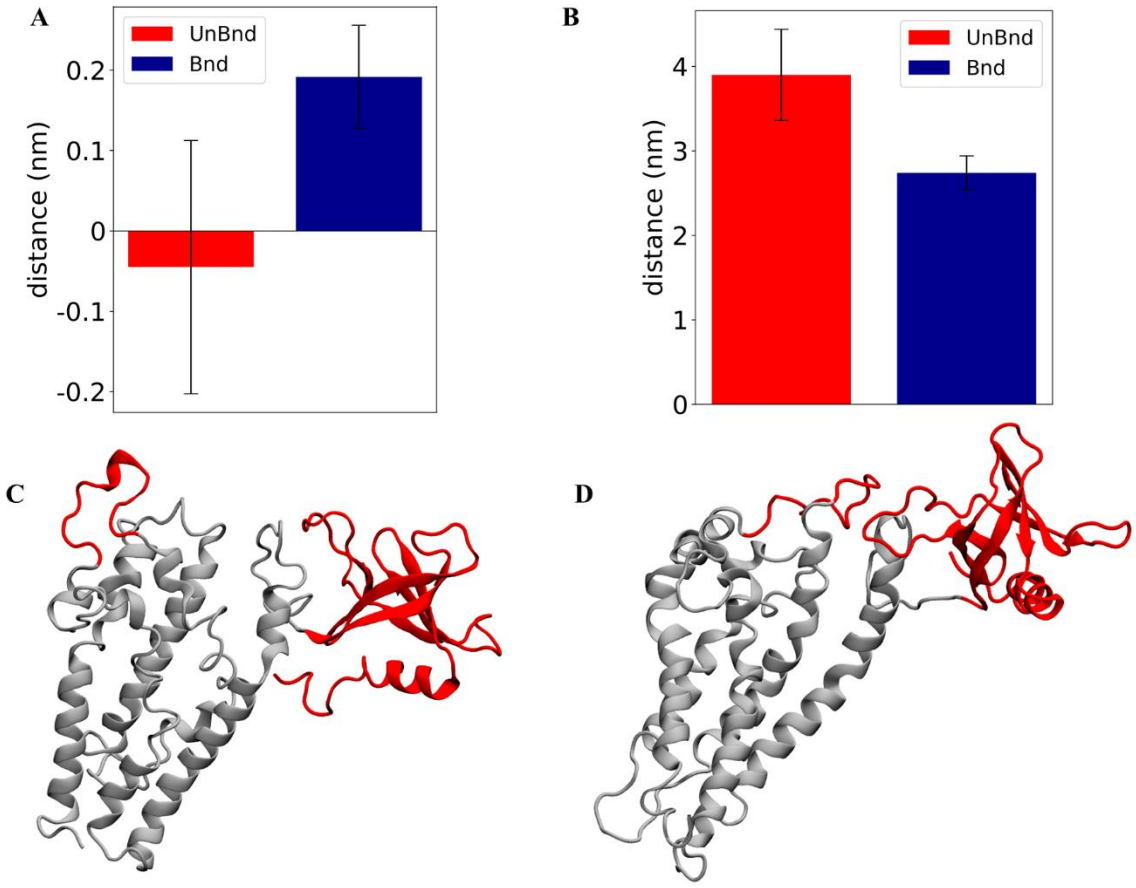

Figure 32. Analysis of the effect of Rac1-α3 interaction. (A) Bar diagram showing the average DH-α3 distances, where the error bars represent the standard deviation of the distribution. (B) Bar diagram showing the average PH-α3-5 distances, where the error bars represent the standard deviation of the distribution. (C) Representative snapshot of Alsin^{UnBnd} where α3-5 and PH are highlighted in red. (D) Representative snapshot of Alsin^{Bnd} where α3-5 and PH are highlighted in red.

To investigate further the role of α3-5 in the stabilization of PH domain position, the same quantities were compared between Alsin^{Rac1} and Alsin^{noRac1} systems. While in the former the last residues of helix α3 were above DH domain centre of mass (0.24 ± 0.04 nm), in the latter they were almost at the same level (0.03 ± 0.04 nm). As a consequence of this displacement, the distance between α3-5 and PH region was slightly increased from 2.83 ± 0.12 nm to 3.19 ± 0.11 nm. It is worth mentioning that, after Rac1 removal, the sole thermal energy was not able to alter the position of PH domain, which remained stuck in an open conformation. Therefore, the increased distance was due to a displacement of α3-5 and helix α3.

5.3.4 Markov State Model of free Alsin

The dynamics of free Alsin was characterized through a MSM built describing the state space in terms of α_{xy} and d_z. From the analysis of the largest implied timescales at increasing lag times, the lag time to build the MSM was set to 9 ns. The number of states determined through

PCCA+ algorithm was set to 5 according to the values of the first 10 implied timescales at 9 ns. The same lag time was used to build the MSM from the 5-state discretization (Figure S6). The model was validated through the Chapman-Kolmogorov test (Figure S7). In Figure 33 are depicted the location of the obtained states in the α_{xy}-d_z plane and a graphical representation of the transition matrix. State 3 corresponded to the region in which the free energy profile of Alsin^{UnBnd} and Alsin^{Bnd} were partially overlapped. It was characterized by a high probability to jump in state 1, while transitions from and to state 2 are less probable. State 1 was quite stable and communicated with all other states, with higher probabilities of jumping to state 2 or 0. In the former the protein was closed on the side of DH domain, while the latter was the least stable state with a high transition probability towards state 4. The last state was quite stable and characterized by the protein being closed from the bottom of DH domain.

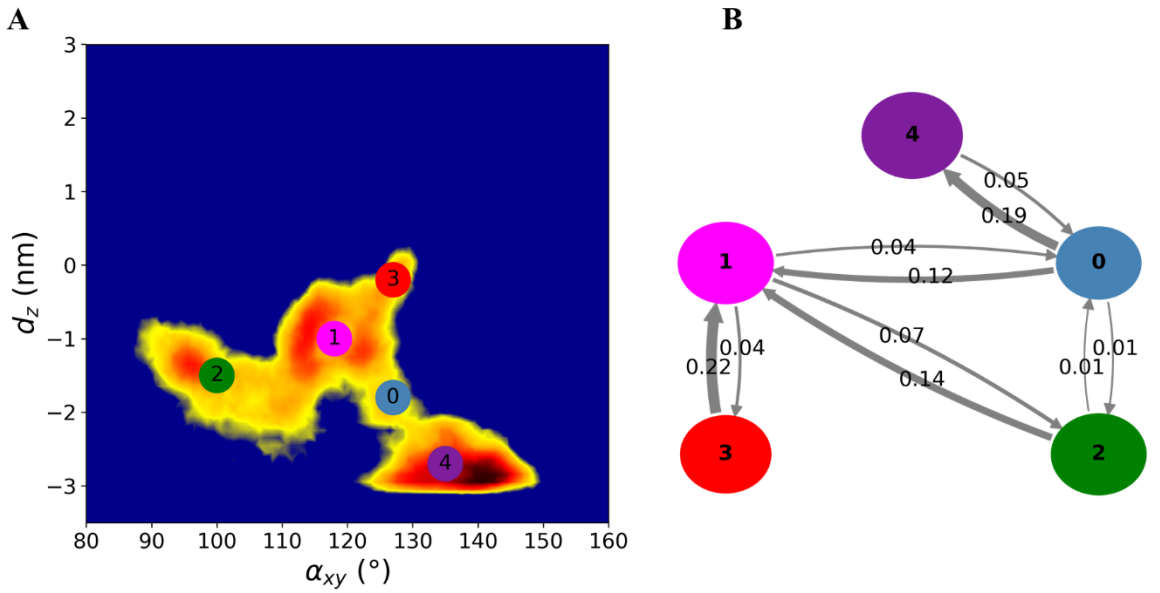

Figure 33. Alsin^{UnBnd} states from MSM analysis. (A) Location of the five states on the α_{xy}-d_z plane. The colours represent the free energy as in Figure 30. (B) Graphical representation of the transition matrix, where transition probabilities were rounded at the second decimal and probabilities lower than 0.005 were not considered. The arrow labels are the jump probabilities between states, the dimension of each sphere is proportional to the self-transition probabilities, and the arrow width is proportional to the probability to observe a jump between the states.

From the analysis of the right eigenvectors, it was possible to identify the four dynamical processes between the states. The slowest one, with an implied timescale of 260 ns, was the transition between states (0, 4) and (1, 2, 3). The second slowest process described the transition between states 2 and (1, 3) and had an implied timescale of 43 ns. The third one was the jump process between states (0, 1), 2, and 3, while the fastest process characterized the transition between states 0, 1, and 3. The latter were associates with 26 ns and 19 ns as timescales,

respectively. A summary of the description of the four dynamical processes is reported in Table 3. Therefore, protein closure hiding the putative dimerization surface was the slowest process, while closure by the side of DH domain was faster.

Table 3. Timescales and transitions that characterize the four dynamical process between the obtained states.

N° process	Transitions	Timescale (ns)
1	$(0,4) \leftrightarrow (1,2,3)$	260
2	$2 \leftrightarrow (1,3)$	43
3	$3 \leftrightarrow (0,1) \leftrightarrow 2$	26
4	$0 \leftrightarrow 1 \leftrightarrow 3$	19

Transition path theory was applied to discover the most probable pathway associated to the transition between the most open conformations, represented by state 3, and the more closed ones, represented by states 2 and 4. The pathway connecting states 3 and 2 was $3 \rightarrow 1 \rightarrow 2$, while the one connecting states 3 and 4 was $3 \rightarrow 1 \rightarrow 0 \rightarrow 4$. States 0 and 1 were also present in the most probable pathway from 2 to 4 and vice versa (Figure 34).

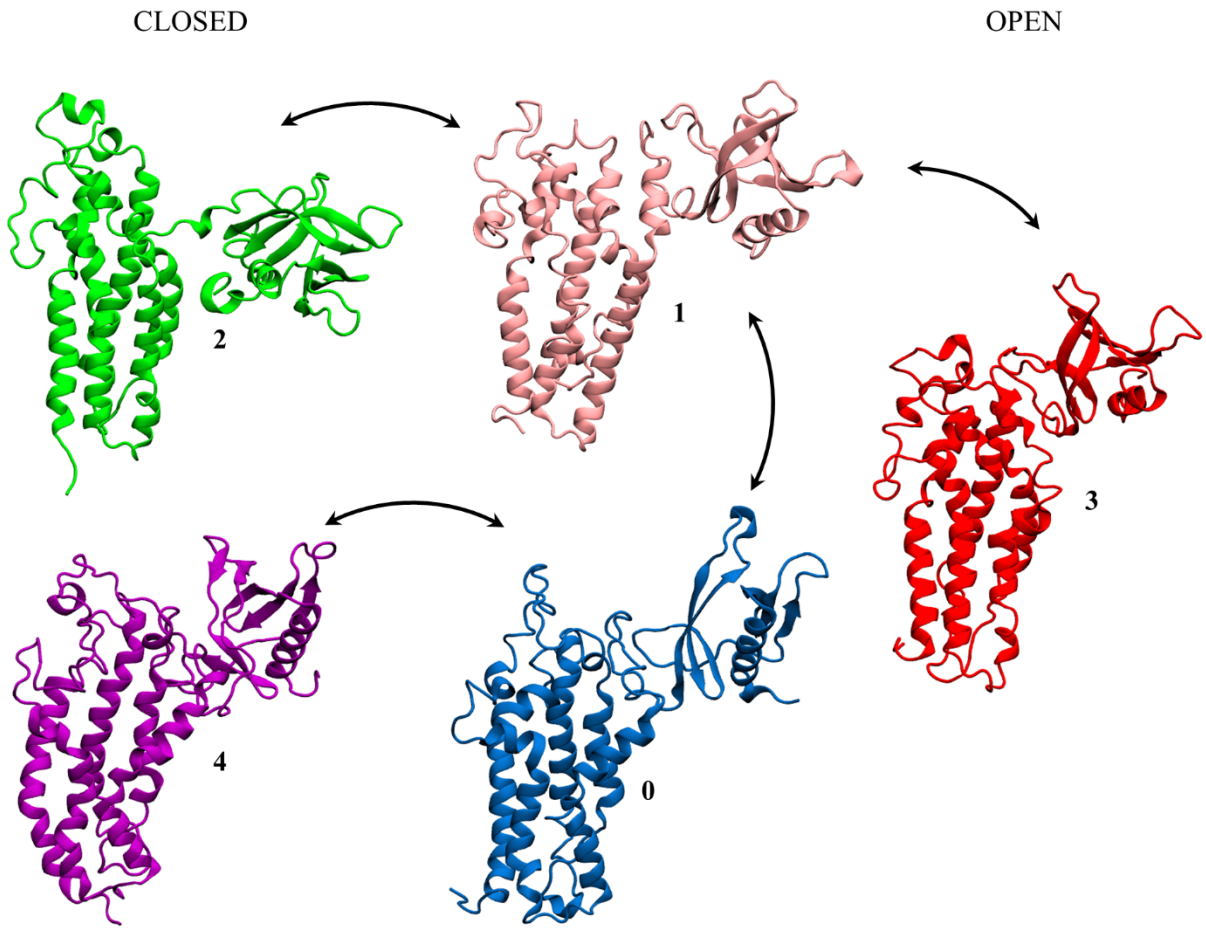

Figure 34. Representation of the most probable transition pathways of Alsin^{UnBnd} between the open to the closed states.

5.4 Discussion

The availability of 3D atomistic models of Alsin is crucial to understand the molecular mechanisms at the basis of its biological functions. Indeed, the first step towards the treatment of ALS2-related pathologies such as IAHSP is a proper comprehension of the protein physiological behaviour. In this work, we focused on Alsin DH/PH domain and developed its first all atom model through I-Tasser suite using 16 templates from RCSB database. The quality of the model was confirmed both by the confidence scores predicted by the employed software, especially the TM-score, and the analysis of the Ramachandran plot (Figure 24). The overall identity scores of the employed templates (Table 2) were generally too low to build a reliable homology model, however it should be considered that they were computed using the complete amino acid sequence while I-Tasser builds the model assembling fragments from different templates.

The alignment of Alsin amino acid sequence with the templates (Figure S4) allowed us to locate the three conserved regions within the first, second, and fifth helices, in agreement with DH/PH domains of other proteins [33], [36], [38], [39]. It should be noticed the presence of a highly conserved histidine in CR2. Indeed, studies in the forefather of DH domains, i.e. the proto-oncogene Dbl, have highlighted the crucial role of an histidine located in the second conserved region in the DH-mediated dimerization [36]. Therefore, H752 in Alsin may play an important role in the Rac1-induced tetramerization.

Since the interaction between Alsin DH domain and Rac1 is known to trigger tetramerization, relocalization at membrane level, and activation of Rab5 through the C-terminal VPS9 domain [4], the effect of such interaction on the dynamics of Alsin DH/PH domain has been studied. Previous analysis on LARG was exploited not only to develop a robust experimental setup but also to identify differences that may be associated with the different role of these two proteins in the GTPase cycle. The main regions involved in Alsin-Rac1 interactions were helices $\alpha 3$, $\alpha 5$, and $\alpha 6$, while lower contact probabilities were found for loops l1, l2, and l3 (Figure 27). Therefore, Rac1 bound also to non structured regions even though they may have only a secondary role. Indeed, while in the second replica of Alsin^{Bnd} the interaction between l3 and Rac1 was direct, in the first trajectory it was mediated by l1. The first conserved region was in contact with the GTPase in the simulations of LARG (see section 4.3.1), but the same evidence was not found in Alsin^{Bnd}, despite the RMSD between the initial configurations was around 2 Å. Since CR1 has been described as crucial for the transforming activity of Rho GEF family of proteins [38], these evidences may explain the different biological function of Alsin. One effect of such interaction was the increased rigidity of the protein within not only two of the regions in contact with Rac1, $\alpha 5$ and $\alpha 6$, but also $\alpha 2$ and the first half of $\alpha 3$ (Figure 28). The latter are exposed on the surface opposite to the one interacting with Rac1, which was found to be involved in the dimerization of other DH domains [36]. Therefore, the mechanical properties of Alsin were locally altered by the interaction with Rac1, stabilizing the putative site of dimerization. Finally, fluctuation in the last residues of helix $\alpha 6$ and in region $\alpha 3$-5 are reduced, suggesting their involvement in the Rac1-driven conformational dynamics (Figure 29).

Previously, it has been proved that Alsin is sequestered in the cytoplasm due to an interaction between RLD and its C-terminus before Rac1 binding [7]. After Rac1 signalling, the protein moves to an open conformation and tetramerizes [4]. The role of Rac1-driven conformational dynamics of DH/PH domain in this signal transduction process has been investigated. The presence of Rac1 stabilized an open and linear conformation where, at most, PH domain could

slightly move above DH centre of mass, near to Rac1 (Figure 30). This movement should not be thought as a transition towards a closed conformation since the presence of Rac1 itself may sterically interfere with the interaction between RLD and Alsin C-terminus. Conversely, in absence of Rac1 different DH-PH relative positions were found and a wider area of the state space was explored. In particular, the regions with low α_{xy} are compatible with an interaction between RLD and C-terminus on the side of DH, while the ones with low d_z may promote an interaction from the bottom of DH, hiding in this way the putative dimerization surface.

The dynamics of Alsin^{UnBnd} was modelled through an MSM to describe the transitions between its different conformations. One open state was found in the region where the PMF of Alsin^{Bnd} and Alsin^{UnBnd} partially overlapped, while two closed states were detected (Figure 33). In one of them, the N- and C-termina of the protein might interact by the side of DH domain, while in the second by the bottom of the domain, hiding the putative dimerization surface. Notably, the latter corresponded with the deepest minimum of the free energy profile. Therefore, in the most stable conformation of Alsin^{UnBnd}, the interaction between RLD and the C-terminus may prevent its tetramerization by hiding the dimerization site within DH domain. The closed and open conformations were connected mainly by one intermediate state, which allowed transitions both between the two closed states and from the open to the closed ones (Figure 34). Thus, due to the sole thermal energy, Alsin DH/PH domain may arrange in a conformation that is more similar to the ones observed in presence of Rac1.

In particular, two main regions within DH domain were characterized by different conformations in Alsin^{UnBnd} and Alsin^{Bnd}. One of them was helix α6, i.e. the one linked to PH domain, whose curvature was higher in absence of Rac1 and, particularly, in the closed states (Figure 31 A, B). Since this helix was involved in the interaction with Rac1, it is possible that its bending is correlated to not only PH motion, but also the propensity of Rac1 and DH domain to bind together. Indeed, it is well known that during their interaction the proteins are not rigid bodies but undergo conformational changes to reach the most favourable arrangement [99]. The second region analysed was α3-5, which was closer to PH domain in Alsin^{Bnd} than Alsin^{UnBnd} (Figure 32 A). Hence, it is possible that an interaction between α3-5 and PH domain plays a role in the stabilization of Alsin open conformation.

To summarize, the present study highlights that free Alsin DH/PH domain exists mainly in a closed state, where the interaction between RLD and C-terminus may hide the putative dimerization surface. In these conformations, the helix α6 is bended such that its interaction

with Rac1 might be unlikely. When, due to thermal energy, PH domain moves to reach a more open state and helix $\alpha 6$ curvature decreases, DH domain may more prone to bind Rac1. Such interaction, together with the one between $\alpha 3$-5 and PH, stabilizes a linear conformation of DH/PH domain, where RLD and the C-terminus should be distanced. In this way, the second conserved region and the C-terminus are exposed and can potentially self-interact, leading to the formation of a tetramer. At the same time, dimerization through DH domain may be favoured by Rac1 interaction since it reduces the fluctuations within the residues that mediate such process.

6 Conclusions

IAHSP has been associated with mutations at the gene encoding for Alsin protein and, to date, there is no cure but it is only possible to treat its manifestations. Indeed, due to the complex nature of this disease and the limited knowledge about the mechanisms leading to a premature neuronal degeneration, the development of specific therapeutic strategies is difficult. Therefore, the understanding of nanoscale phenomena involved in Alsin-mediated physiological and pathological pathways is crucial to speed up the process of drug discovery. In this framework, the first step is represented by the development of Alsin 3D structure and the analysis of its dynamics.

This M.Sc. thesis focuses on Alsin DH/PH domain and its binding partner Rac1, since their interaction is the first event within Alsin-mediated pathways. Due to the lack of an experimental molecular structure, the first goal was the construction of a high-quality homology model. Then, to study the dynamics of this domain and obtain robust results, the employed experimental setup was tailored and validated reproducing previous findings from literature on a homologous domains. The analysis on Alsin DH/PH region, both with and without Rac1, revealed that its CR1 is not involved in the interaction with its binding partner, unlike other Rho GEF proteins. This evidence suggests the molecular basis of its different biological function. Moreover, the mechanism by which Rac1 may trigger the dimerization through DH region has been investigated. While stabilizing a more open and linear conformation of the whole domain by changing the DH-PH relative position, the interaction with Rac1 caused a local increase in the mechanical properties within the putative dimerization site. Therefore, the presence of Rac1 moves away Alsin N- and C-termina, exposes the dimerization site, and reduces the fluctuation of residues within this region. At the same time, the dynamics of free Alsin was analysed through a Markov State Model, suggesting in this way the possible pathways linking the closed and open states of the domain.

These results represent an important starting point for further analysis, such as the study of the dimerization both using computational and experimental methodologies. Moreover, different computational analysis on the developed model might provide useful information to experimentally resolve Alsin structure.

7 Acknowledgements

I would like to show my gratitude to my supervisor Prof. Marco Agostino Deriu for having given me the opportunity to work on this thesis. His positivity, enthusiasm, and professionality constantly increased my interest not only in molecular modelling. I am extremely thankful to Marcello Miceli for his precious advice and the critical evaluation of my work. His constant support was determining in increasing the quality of this thesis.

I would like to thank Olivia's family, the Help Olly association, and the project Seed Grant Spring 2020 – IAHSP, since their effort to shed light into IAHSP allowed me to get in touch with this topic which captured my interest. I would thank the Telethon foundation, that gives the opportunity to study rare and complex pathologies and train students and researchers.

I would like to thank Vanessa, Alice, and Sara, whom I encountered from the first years at Politecnico. Thank you, Davide, Barbara, Fabio, Luca, Beatrice, and all my other friends for having shared with me these last years, from small things to future ambitions. Thanks to Riccardo for the precious exchange of views and having worked together with me during the thesis. Thanks to my family, who has supported me throughout the university studies, for always being here. Finally, I cannot express my gratitude to Silvia, my life partner, who challenges me to be my best self.

Marco Cannariato

8 Supplementary information

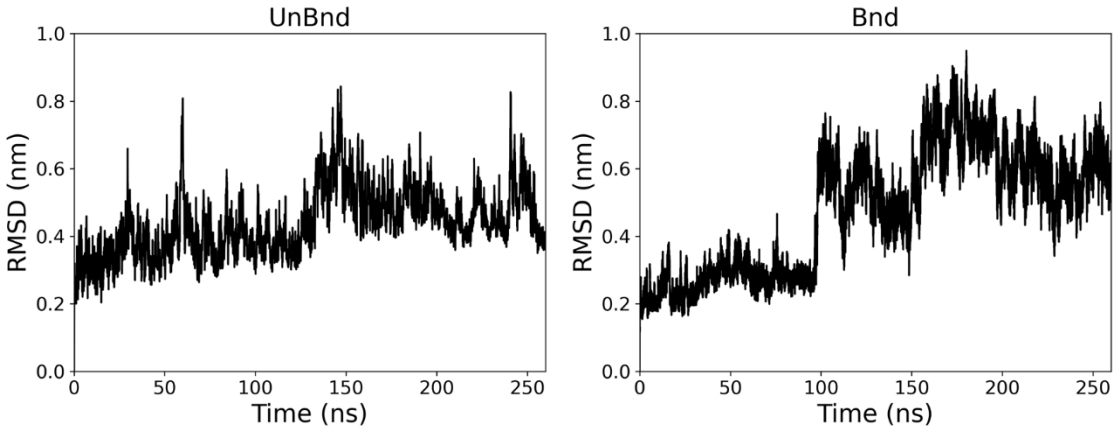

Figure S1. RMSD plots for the UnBnd and Bnd states of LARG.

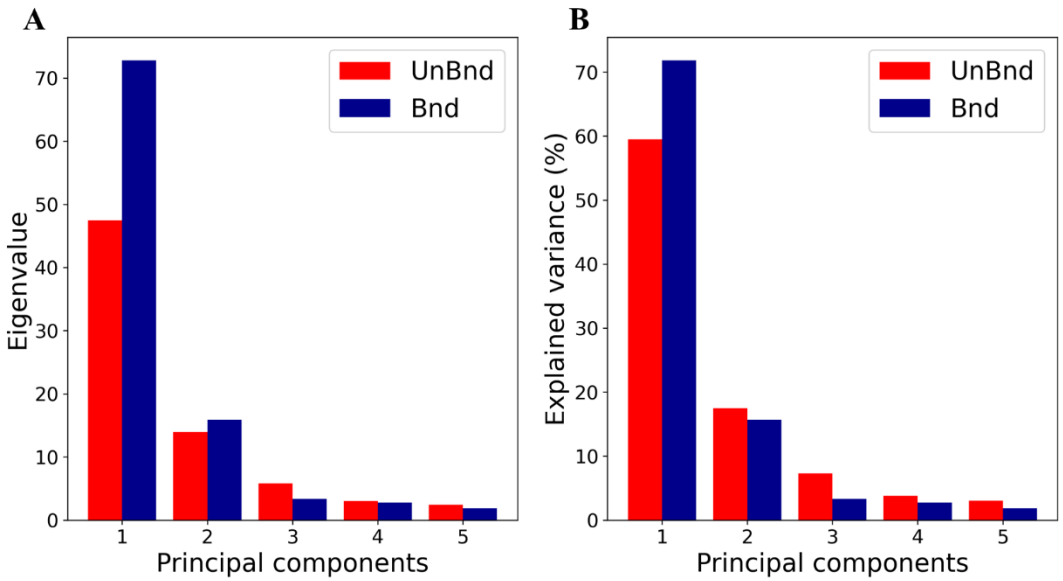

Figure S2. (A) Comparison of the eigenvalues of the covariance matrix. (B) Percentage of the total variance explained by the principal components.

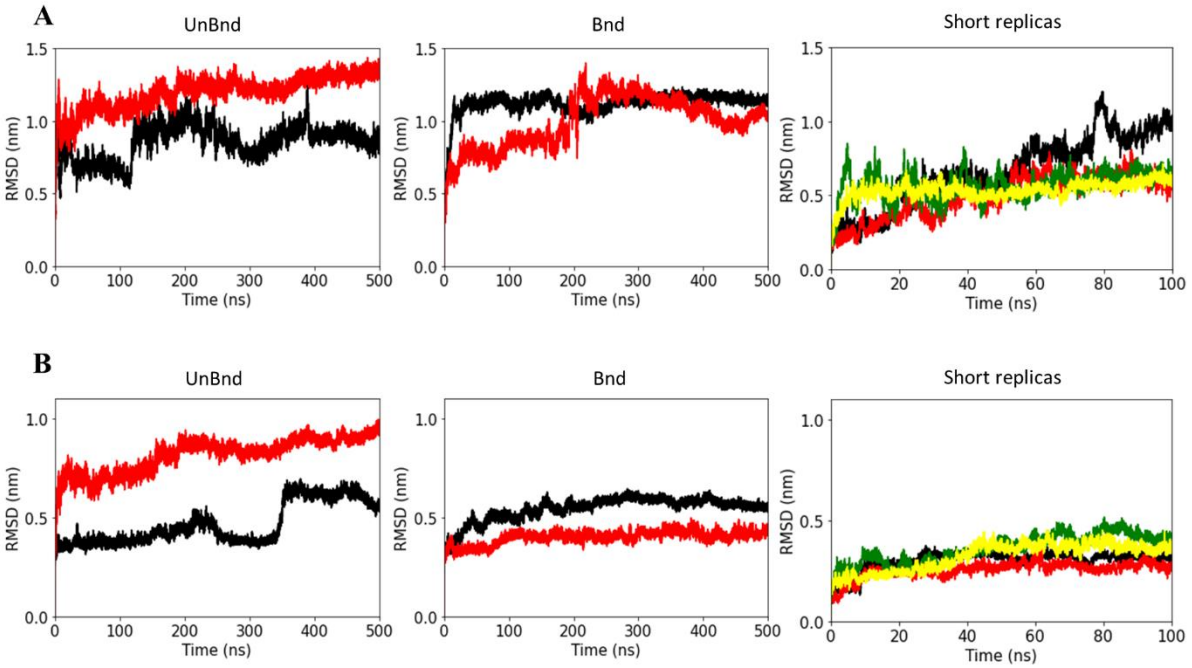

Figure S3. RMSD plots of (A) protein C-alphas and (B) DH domain C-alphas from the initial configuration.

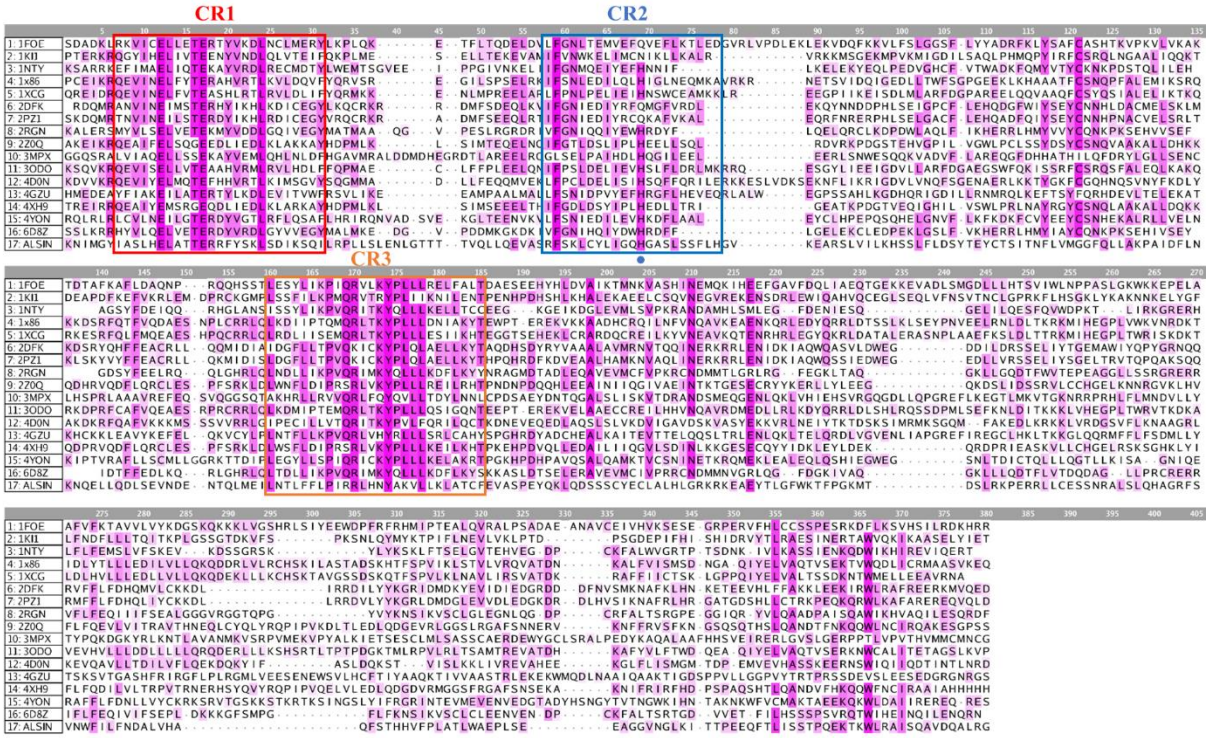

Figure S4. Alignment of Alsin and the 16 templates used by I-Tasser to build the homology model. The templates residues that were not aligned with Alsin sequence are not included. Residues are coloured according to the identity percentage over the 17 sequences. The three conserved regions are highlighted in red, blue, and orange.

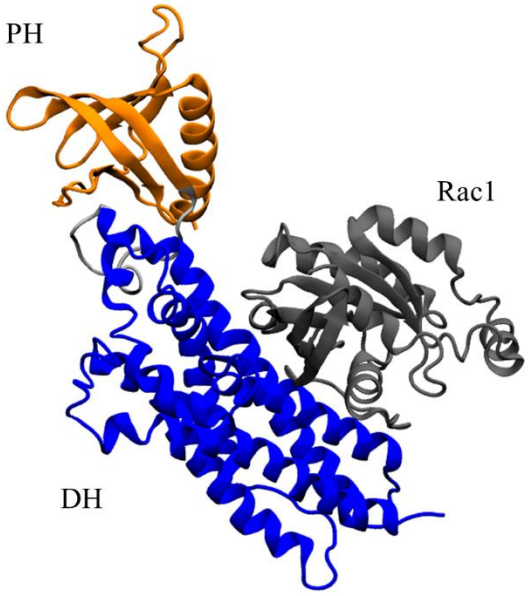

Figure S5. Initial configuration of Alsin^{Bnd}, obtained superimposing Alsin and Rac1 to LARG and RhoA, respectively. DH domain, PH domain, and Rac1 are represented in blue, orange, and grey, respectively.

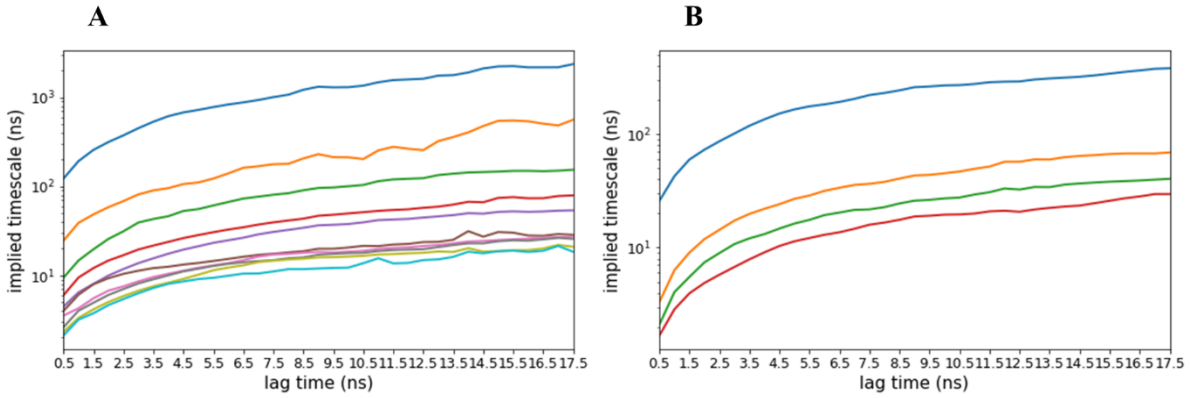

Figure S6. Implied timescales plot using the discretization in (A) 1000 clusters from K-Centers and (B) 5 states obtained through PCCA+.

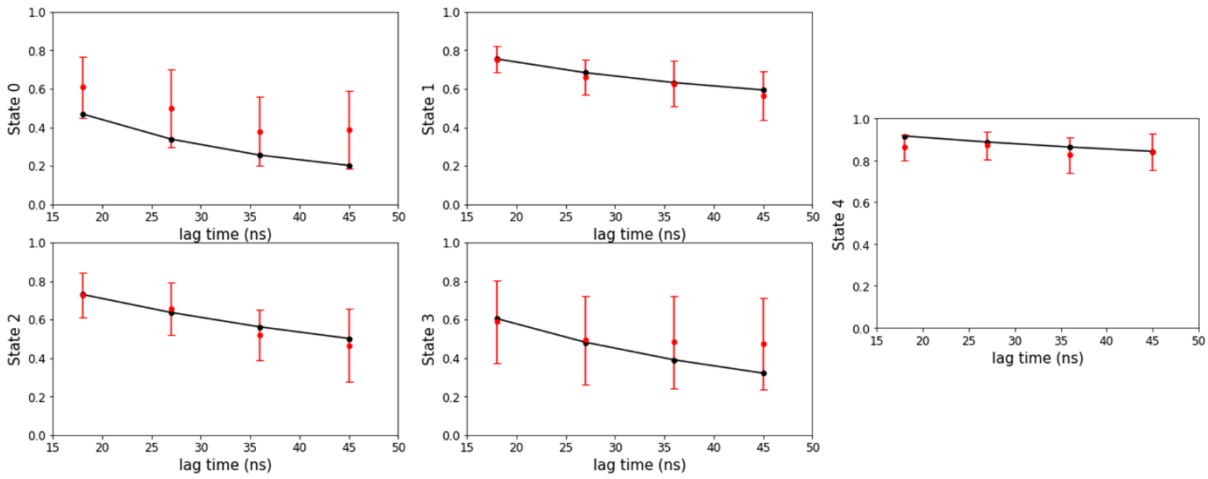

Figure S7. Chapman-Kolmogorov test of the Markov State Model. Only self-transition probabilities are represented. The black line represents the transition probability estimated propagating the MSM, red points and error bars represent the mean and standard deviation of the transition probability at multiples of the lag time estimates through the Bayesian approach.

9 References

[1] C. Blackstone, “Hereditary spastic paraplegia,” 2018, pp. 633–652.

[2] J. K. Fink, “Hereditary spastic paraplegia: clinico-pathologic features and emerging molecular mechanisms,” *Acta Neuropathol.*, vol. 126, no. 3, pp. 307–328, 2013, doi: 10.1007/s00401-013-1115-8.Hereditary.

[3] R. W. Orrell, *ALS2-Related Disorder*. 1993.

[4] K. Sato *et al.*, “Altered oligomeric states in pathogenic ALS2 variants associated with juvenile motor neuron diseases cause loss of ALS2-mediated endosomal function,” *J. Biol. Chem.*, vol. 293, no. 44, pp. 17135–17153, 2018, doi: 10.1074/jbc.RA118.003849.

[5] A. Otomo, “ALS2, a novel guanine nucleotide exchange factor for the small GTPase Rab5, is implicated in endosomal dynamics,” *Hum. Mol. Genet.*, vol. 12, no. 14, pp. 1671–1687, Jul. 2003, doi: 10.1093/hmg/ddg184.

[6] S. Hadano *et al.*, “Mice deficient in the Rab5 guanine nucleotide exchange factor ALS2/alsin exhibit age-dependent neurological deficits and altered endosome trafficking,” *Hum. Mol. Genet.*, vol. 15, no. 2, pp. 233–250, Jan. 2006, doi: 10.1093/hmg/ddi440.

[7] R. Kunita, A. Otomo, H. Mizumura, K. Suzuki-Utsunomiya, S. Hadano, and J. E. Ikeda, “The Rab5 activator ALS2/alsin acts as a novel Rac1 effector through Rac1-activated endocytosis,” *J. Biol. Chem.*, vol. 282, no. 22, pp. 16599–16611, 2007, doi: 10.1074/jbc.M610682200.

[8] M. Karplus and J. A. McCammon, “Molecular dynamics simulations of biomolecules,” *Nat. Struct. Biol.*, vol. 9, no. 9, pp. 646–652, Sep. 2002, doi: 10.1038/nsb0902-646.

[9] M. Karplus and J. Kuriyan, “Molecular dynamics and protein function,” *Proc. Natl. Acad. Sci.*, vol. 102, no. 19, pp. 6679–6685, May 2005, doi: 10.1073/pnas.0408930102.

[10] M. Helal *et al.*, “Clinical presentation and natural history of infantile-onset ascending spastic paralysis from three families with an ALS2 founder variant,” *Neurol. Sci.*, vol. 39, no. 11, pp. 1917–1925, Nov. 2018, doi: 10.1007/s10072-018-3526-8.

[11] E. Eymard-Pierre *et al.*, “Infantile-Onset Ascending Hereditary Spastic Paralysis Is Associated with Mutations in the Alsin Gene,” *Am. J. Hum. Genet.*, vol. 71, no. 3, pp.

518–527, Sep. 2002, doi: 10.1086/342359.

[12] M. Gautam, J. H. Jara, G. Sekerkova, M. V. Yasvoina, M. Martina, and P. H. Özdinler, "Absence of alsin function leads to corticospinal motor neuron vulnerability via novel disease mechanisms," *Hum. Mol. Genet.*, vol. 25, no. 6, pp. 1074–1087, 2015, doi: 10.1093/hmg/ddv631.

[13] R. Sprute *et al.*, "Genotype–phenotype correlation in seven motor neuron disease families with novel <scp> *ALS2* </scp> mutations," *Am. J. Med. Genet. Part A*, p. ajmg.a.61951, Nov. 2020, doi: 10.1002/ajmg.a.61951.

[14] K. Wennerberg, K. L. Rossman, and C. J. Der, "The Ras superfamily at a glance," *J. Cell Sci.*, vol. 118, no. 5, pp. 843–846, 2005, doi: 10.1242/jcs.01660.

[15] K. H. Pedone, C. J. Der, and V. Kitainda, "Small GTPases," in *Reference Module in Life Sciences*, Elsevier, 2021.

[16] A. Kumar, V. Rajendran, R. Sethumadhavan, and R. Purohit, "Molecular Dynamic Simulation Reveals Damaging Impact of RAC1 F28L Mutation in the Switch I Region," *PLoS One*, vol. 8, no. 10, p. e77453, Oct. 2013, doi: 10.1371/journal.pone.0077453.

[17] D. C. Soares, P. N. Barlow, D. J. Porteous, and R. S. Devon, "An interrupted beta-propeller and protein disorder: Structural bioinformatics insights into the N-terminus of alsin," *J. Mol. Model.*, vol. 15, no. 2, pp. 113–122, 2009, doi: 10.1007/s00894-008-0381-1.

[18] K. Kanekura, Y. Hashimoto, T. Niikura, S. Aiso, M. Matsuoka, and I. Nishimoto, "Alsin, the Product of ALS2 Gene, Suppresses SOD1 Mutant Neurotoxicity through RhoGEF Domain by Interacting with SOD1 Mutants," *J. Biol. Chem.*, vol. 279, no. 18, pp. 19247–19256, 2004, doi: 10.1074/jbc.M313236200.

[19] I. Mellman, "ENDOCYTOSIS AND MOLECULAR SORTING," *Annu. Rev. Cell Dev. Biol.*, vol. 12, no. 1, pp. 575–625, Nov. 1996, doi: 10.1146/annurev.cellbio.12.1.575.

[20] J. Rink, E. Ghigo, Y. Kalaidzidis, and M. Zerial, "Rab Conversion as a Mechanism of Progression from Early to Late Endosomes," *Cell*, vol. 122, no. 5, pp. 735–749, Sep. 2005, doi: 10.1016/j.cell.2005.06.043.

[21] E. L. Racoosin and J. A. Swanson, "Macropinosome maturation and fusion with tubular lysosomes in macrophages.," *J. Cell Biol.*, vol. 121, no. 5, pp. 1011–1020, Jun. 1993,

doi: 10.1083/jcb.121.5.1011.

[22] S. Hadano *et al.*, “Loss of ALS2/Alsin exacerbates motor dysfunction in a SOD1H46R-expressing mouse ALS model by disturbing endolysosomal trafficking,” *PLoS One*, vol. 5, no. 3, 2010, doi: 10.1371/journal.pone.0009805.

[23] A. Otomo, L. Pan, and S. Hadano, “Dysregulation of the autophagy-endolysosomal system in amyotrophic lateral sclerosis and related motor neuron diseases,” *Neurol. Res. Int.*, vol. 2012, 2012, doi: 10.1155/2012/498428.

[24] A. Otomo, R. Kunita, K. Suzuki-utsunomiya, H. Mizumura, and K. Onoe, “Biochemical and Biophysical Research Communications ALS2 / alsin deficiency in neurons leads to mild defects in macropinocytosis and axonal growth,” vol. 370, pp. 87–92, 2008, doi: 10.1016/j.bbrc.2008.01.177.

[25] R. Kunita *et al.*, “Homo-oligomerization of ALS2 through its unique carboxyl-terminal regions is essential for the ALS2-associated Rab5 guanine nucleotide exchange activity and its regulatory function on endosome trafficking,” *J. Biol. Chem.*, vol. 279, no. 37, pp. 38626–38635, 2004, doi: 10.1074/jbc.M406120200.

[26] A. Otomo, R. Kunita, K. Suzuki-Utsunomiya, J. E. Ikeda, and S. Hadano, “Defective relocalization of ALS2/alsin missense mutants to Rac1-induced macropinosomes accounts for loss of their cellular function and leads to disturbed amphisome formation,” *FEBS Lett.*, vol. 585, no. 5, pp. 730–736, 2011, doi: 10.1016/j.febslet.2011.01.045.

[27] F. S. Hsu, S. Spannl, C. Ferguson, A. A. Hyman, R. G. Parton, and M. Zerial, “Rab5 and Alsin regulate stress-activated cytoprotective signaling on mitochondria,” *Elife*, vol. 7, pp. 1–37, 2018, doi: 10.7554/eLife.32282.

[28] Q. Li, N. Y. Spencer, N. J. Pantazis, and J. F. Engelhardt, “Alsin and SOD1 G93A proteins regulate endosomal reactive oxygen species production by glial cells and proinflammatory pathways responsible for neurotoxicity,” *J. Biol. Chem.*, vol. 286, no. 46, pp. 40151–40162, 2011, doi: 10.1074/jbc.M111.279711.

[29] H. Cai, “Loss of ALS2 Function Is Insufficient to Trigger Motor Neuron Degeneration in Knock-Out Mice But Predisposes Neurons to Oxidative Stress,” *J. Neurosci.*, vol. 25, no. 33, pp. 7567–7574, Aug. 2005, doi: 10.1523/JNEUROSCI.1645-05.2005.

[30] K. Forsberg *et al.*, “Misfolded SOD1 inclusions in patients with mutations in C9orf72

and other ALS/FTD-associated genes," *J. Neurol. Neurosurg. Psychiatry*, pp. 1–9, 2019, doi: 10.1136/jnnp-2018-319386.

[31] K. Deinhardt *et al.*, "Rab5 and Rab7 Control Endocytic Sorting along the Axonal Retrograde Transport Pathway," *Neuron*, vol. 52, no. 2, pp. 293–305, 2006, doi: 10.1016/j.neuron.2006.08.018.

[32] S. Niftullayev and N. Lamarche-Vane, "Regulators of Rho GTPases in the Nervous System: Molecular Implication in Axon Guidance and Neurological Disorders," *Int. J. Mol. Sci.*, vol. 20, no. 6, p. 1497, Mar. 2019, doi: 10.3390/ijms20061497.

[33] B. Aghazadeh *et al.*, "Structure and mutagenesis of the Dbl homology domain," *Nat. Struct. Biol.*, vol. 5, no. 12, pp. 1098–1107, Dec. 1998, doi: 10.1038/4209.

[34] A. Felline *et al.*, "Interconnecting flexibility , structural communication , and function in RhoGEF oncoproteins," 2019, doi: 10.1021/acs.jcim.9b00271.

[35] C. M. Lucato *et al.*, "The Phosphatidylinositol (3,4,5)-Trisphosphate-dependent Rac Exchanger 1·Ras-related C3 Botulinum Toxin Substrate 1 (P-Rex1·Rac1) Complex Reveals the Basis of Rac1 Activation in Breast Cancer Cells," *J. Biol. Chem.*, vol. 290, no. 34, pp. 20827–20840, Aug. 2015, doi: 10.1074/jbc.M115.660456.

[36] Y. Z. Kejin, D. Balazs, B. Feng, and Zheng, "Oligomerization of DH Domain Is Essential for Dbl-Induced Transformation," *Mol. Cell. Biol.*, vol. 21, no. 2, pp. 425–437, Jan. 2001, doi: 10.1128/MCB.21.2.425-437.2001.

[37] S. M. Soisson, A. S. Nimnual, M. Uy, D. Bar-Sagi, and J. Kuriyan, "Crystal Structure of the Dbl and Pleckstrin Homology Domains from the Human Son of Sevenless Protein," *Cell*, vol. 95, no. 2, pp. 259–268, Oct. 1998, doi: 10.1016/S0092-8674(00)81756-0.

[38] K. Zhu, B. Debreceni, R. Li, and Y. Zheng, "Identification of Rho GTPase-dependent Sites in the Dbl Homology Domain of Oncogenic Dbl That Are Required for Transformation," *J. Biol. Chem.*, vol. 275, no. 34, pp. 25993–26001, Aug. 2000, doi: 10.1074/jbc.M003780200.

[39] D. K. Worthylake, K. L. Rossman, and J. Sondek, "Crystal structure of Rac1 in complex with the guanine nucleotide exchange region of Tiam1," *Nature*, vol. 408, no. 6813, pp. 682–688, Dec. 2000, doi: 10.1038/35047014.

[40] K. Verhoeven *et al.*, "Slowed Conduction and Thin Myelination of Peripheral Nerves

Associated with Mutant Rho Guanine-Nucleotide Exchange Factor 10," *Am. J. Hum. Genet.*, vol. 73, no. 4, pp. 926–932, Oct. 2003, doi: 10.1086/378159.

[41] M. Chen *et al.*, "Structure and regulation of human epithelial cell transforming 2 protein," *Proc. Natl. Acad. Sci.*, vol. 117, no. 2, pp. 1027–1035, Jan. 2020, doi: 10.1073/pnas.1913054117.

[42] X. He, Y.-C. Kuo, T. J. Rosche, and X. Zhang, "Structural Basis for Autoinhibition of the Guanine Nucleotide Exchange Factor FARP2," *Structure*, vol. 21, no. 3, pp. 355–364, Mar. 2013, doi: 10.1016/j.str.2013.01.001.

[43] H. S. Carr, C. A. Morris, S. Menon, E. H. Song, and J. A. Frost, "Rac1 Controls the Subcellular Localization of the Rho Guanine Nucleotide Exchange Factor Net1A To Regulate Focal Adhesion Formation and Cell Spreading," *Mol. Cell. Biol.*, vol. 33, no. 3, pp. 622–634, Feb. 2013, doi: 10.1128/MCB.00980-12.

[44] M. A. M. Frasa *et al.*, "Armus Is a Rac1 Effector that Inactivates Rab7 and Regulates E-Cadherin Degradation," *Curr. Biol.*, vol. 20, no. 3, pp. 198–208, Feb. 2010, doi: 10.1016/j.cub.2009.12.053.

[45] X. Pang *et al.*, "A PH Domain in ACAP1 Possesses Key Features of the BAR Domain in Promoting Membrane Curvature," *Dev. Cell*, vol. 31, no. 1, pp. 73–86, Oct. 2014, doi: 10.1016/j.devcel.2014.08.020.

[46] S.-A. Kim, P. O. Vacratsis, R. Firestein, M. L. Cleary, and J. E. Dixon, "Regulation of myotubularin-related (MTMR)2 phosphatidylinositol phosphatase by MTMR5, a catalytically inactive phosphatase," *Proc. Natl. Acad. Sci.*, vol. 100, no. 8, pp. 4492–4497, Apr. 2003, doi: 10.1073/pnas.0431052100.

[47] N. Umeki, H. S. Jung, T. Sakai, O. Sato, R. Ikebe, and M. Ikebe, "Phospholipid-dependent regulation of the motor activity of myosin X," *Nat. Struct. Mol. Biol.*, vol. 18, no. 7, pp. 783–788, Jul. 2011, doi: 10.1038/nsmb.2065.

[48] I. Hers, E. E. Vincent, and J. M. Tavaré, "Akt signalling in health and disease ☆," *Cell. Signal.*, vol. 23, no. 10, pp. 1515–1527, 2011, doi: 10.1016/j.cellsig.2011.05.004.

[49] K. D. Swanson *et al.*, "Article The Skap-hom Dimerization and PH Domains Comprise a 3 0 -Phosphoinositide-Gated Molecular Switch," *Mol. Cell*, vol. 32, no. 4, pp. 564–575, 2008, doi: 10.1016/j.molcel.2008.09.022.

[50] M. A. Baumeister, K. L. Rossman, J. Sondek, and M. A. Lemmon, “The Dbs PH domain contributes independently to membrane targeting and regulation of guanine nucleotide-exchange activity,” *Biochem. J.*, vol. 400, no. 3, pp. 563–572, Dec. 2006, doi: 10.1042/BJ20061020.

[51] J. M. Berg, J. L. Tymoczko, and L. Stryer, “Protein Structure and Function,” in *Biochemistry. 5th edition*, New York, New York, USA, 2002.

[52] F. Raimondi, A. Felline, and F. Fanelli, “Catching Functional Modes and Structural Communication in Dbl Family Rho Guanine Nucleotide Exchange Factors,” *J. Chem. Inf. Model.*, vol. 55, no. 9, pp. 1878–1893, Sep. 2015, doi: 10.1021/acs.jcim.5b00122.

[53] Y. C. Martin *et al.*, “Glossary of terms used in computational drug design, part II (IUPAC Recommendations 2015),” *Pure Appl. Chem.*, vol. 88, no. 3, pp. 239–264, Mar. 2016, doi: 10.1515/pac-2012-1204.

[54] J. Chapman, “Improving the Functional Control of Ferroelectrics using Insights from Atomistic Modelling,” 2018.

[55] J. Westbrook, “The Protein Data Bank: unifying the archive,” *Nucleic Acids Res.*, vol. 30, no. 1, pp. 245–248, Jan. 2002, doi: 10.1093/nar/30.1.245.

[56] C. Chothia and A. M. Lesk, “The relation between the divergence of sequence and structure in proteins.,” *EMBO J.*, vol. 5, no. 4, pp. 823–6, Apr. 1986, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/3709526.

[57] A. Roy, A. Kucukural, and Y. Zhang, “I-TASSER: a unified platform for automated protein structure and function prediction,” *Nat. Protoc.*, vol. 5, no. 4, pp. 725–738, Apr. 2010, doi: 10.1038/nprot.2010.5.

[58] J. Zhang, Q. Bai, H. Pérez-Sánchez, S. Shang, X. An, and X. Yao, “Investigation of ECD conformational transition mechanism of GLP-1R by molecular dynamics simulations and Markov state model,” *Phys. Chem. Chem. Phys.*, vol. 21, no. 16, pp. 8470–8481, 2019, doi: 10.1039/C9CP00080A.

[59] J.-H. Prinz *et al.*, “Markov models of molecular kinetics: Generation and validation,” *J. Chem. Phys.*, vol. 134, no. 17, p. 174105, May 2011, doi: 10.1063/1.3565032.

[60] F. Noé, I. Horenko, C. Schütte, and J. C. Smith, “Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states,” *J. Chem. Phys.*,

vol. 126, no. 15, p. 155102, Apr. 2007, doi: 10.1063/1.2714539.

[61] P. Metzner, F. Noé, and C. Schütte, "Estimating the sampling error: Distribution of transition matrices and functions of transition matrices for given trajectory data," *Phys. Rev. E*, vol. 80, no. 2, p. 021106, Aug. 2009, doi: 10.1103/PhysRevE.80.021106.

[62] W. Wang, S. Cao, L. Zhu, and X. Huang, "Constructing Markov State Models to elucidate the functional conformational changes of complex biomolecules," *WIREs Comput. Mol. Sci.*, vol. 8, no. 1, Jan. 2018, doi: 10.1002/wcms.1343.

[63] P. Metzner, C. Schütte, and E. Vanden-Eijnden, "Transition Path Theory for Markov Jump Processes," *Multiscale Model. Simul.*, vol. 7, no. 3, pp. 1192–1219, Jan. 2009, doi: 10.1137/070699500.

[64] P. J. Kourlas *et al.*, "Identification of a gene at 11q23 encoding a guanine nucleotide exchange factor: Evidence for its fusion with MLL in acute myeloid leukemia," *Proc. Natl. Acad. Sci.*, vol. 97, no. 5, pp. 2145–2150, Feb. 2000, doi: 10.1073/pnas.040569197.

[65] M. A. Booden, D. P. Siderovski, and C. J. Der, "Leukemia-Associated Rho Guanine Nucleotide Exchange Factor Promotes Gαq-Coupled Activation of RhoA," *Mol. Cell. Biol.*, vol. 22, no. 12, pp. 4053–4061, Jun. 2002, doi: 10.1128/MCB.22.12.4053-4061.2002.

[66] D. C. T. Ong *et al.*, "LARG at chromosome 11q23 has functional characteristics of a tumor suppressor in human breast and colorectal cancer," *Oncogene*, vol. 28, no. 47, pp. 4189–4200, Nov. 2009, doi: 10.1038/onc.2009.266.

[67] R. Kristelly, G. Gao, and J. J. G. Tesmer, "Structural Determinants of RhoA Binding and Nucleotide Exchange in Leukemia-associated Rho Guanine-Nucleotide Exchange Factor," *J. Biol. Chem.*, vol. 279, no. 45, pp. 47352–47362, Nov. 2004, doi: 10.1074/jbc.M406056200.

[68] Lindahl, Abraham, Hess, and van der Spoel, "GROMACS 2020.4 Source code," Oct. 2020, doi: 10.5281/ZENODO.4054979.

[69] K. Lindorff-Larsen *et al.*, "Improved side-chain torsion potentials for the Amber ff99SB protein force field," *Proteins Struct. Funct. Bioinforma.*, vol. 78, no. 8, pp. 1950–1958, Jun. 2010, doi: 10.1002/prot.22711.

[70] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein,

"Comparison of simple potential functions for simulating liquid water," *J. Chem. Phys.*, vol. 79, no. 2, pp. 926–935, Jul. 1983, doi: 10.1063/1.445869.

[71] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," *J. Chem. Phys.*, vol. 81, no. 8, pp. 3684–3690, Oct. 1984, doi: 10.1063/1.448118.

[72] W. Humphrey, A. Dalke, and K. Schulten, "VMD - Visual Molecular Dynamics," *J. Mol. Graph. Model.*, vol. 14, pp. 33–38, 1996.

[73] M. A. Deriu *et al.*, "Investigation of the Josephin Domain Protein-Protein Interaction by Molecular Dynamics," *PLoS One*, vol. 9, no. 9, p. e108677, Sep. 2014, doi: 10.1371/journal.pone.0108677.

[74] I. Navizet, F. Cailliez, and R. Lavery, "Probing Protein Mechanics: Residue-Level Properties and Their Use in Defining Domains," *Biophys. J.*, vol. 87, no. 3, pp. 1426–1435, Sep. 2004, doi: 10.1529/biophysj.104.042085.

[75] R. Lavery and S. Sacquin-Mora, "Protein mechanics: a route from structure to function," *J. Biosci.*, vol. 32, no. S1, pp. 891–898, Aug. 2007, doi: 10.1007/s12038-007-0089-x.

[76] R. Gowers *et al.*, "MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations," 2016, pp. 98–105, doi: 10.25080/Majora-629e541a-00e.

[77] N. Michaud-Agrawal, E. J. Denning, T. B. Woolf, and O. Beckstein, "MDAnalysis: A toolkit for the analysis of molecular dynamics simulations," *J. Comput. Chem.*, vol. 32, no. 10, pp. 2319–2327, Jul. 2011, doi: 10.1002/jcc.21787.

[78] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, 2007, doi: 10.1109/MCSE.2007.55.

[79] K. R. Acharya and M. D. Lloyd, "The advantages and limitations of protein crystal structures," *Trends Pharmacol. Sci.*, vol. 26, no. 1, pp. 10–14, Jan. 2005, doi: 10.1016/j.tips.2004.10.011.

[80] Y. Zhang, "I-TASSER server for protein 3D structure prediction," *BMC Bioinformatics*, vol. 9, no. 1, p. 40, Dec. 2008, doi: 10.1186/1471-2105-9-40.

[81] J. Yang, A. Roy, and Y. Zhang, "Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile

alignment," *Bioinformatics*, vol. 29, no. 20, pp. 2588–2595, Oct. 2013, doi: 10.1093/bioinformatics/btt447.

[82] C. C. G. ULC, "Molecular Operating Environment (MOE)." Montreal, QC, Canada, H3A 2R7, 2019.

[83] M. Krauthammer *et al.*, "Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma," *Nat. Genet.*, vol. 44, no. 9, pp. 1006–1014, Sep. 2012, doi: 10.1038/ng.2359.

[84] R. Ribarics, M. Kenn, R. Karch, N. Ilieva, and W. Schreiner, "Geometry Dynamics of α -Helices in Different Class I Major Histocompatibility Complexes," *J. Immunol. Res.*, vol. 2015, pp. 1–20, 2015, doi: 10.1155/2015/173593.

[85] K. A. Beauchamp, G. R. Bowman, T. J. Lane, L. Maibaum, I. S. Haque, and V. S. Pande, "MSMBuilder2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale," *J. Chem. Theory Comput.*, vol. 7, no. 10, pp. 3412–3419, Oct. 2011, doi: 10.1021/ct200463m.

[86] M. K. Scherer *et al.*, "PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models," *J. Chem. Theory Comput.*, vol. 11, no. 11, pp. 5525–5542, Nov. 2015, doi: 10.1021/acs.jctc.5b00743.

[87] J. T. Snyder *et al.*, "Structural basis for the selective activation of Rho GTPases by Dbl exchange factors," *Nat. Struct. Biol.*, vol. 9, no. 6, pp. 468–475, Jun. 2002, doi: 10.1038/nsb796.

[88] K. R. Skowronek, F. Guo, Y. Zheng, and N. Nassar, "The C-terminal Basic Tail of RhoG Assists the Guanine Nucleotide Exchange Factor Trio in Binding to Phospholipids," *J. Biol. Chem.*, vol. 279, no. 36, pp. 37895–37907, Sep. 2004, doi: 10.1074/jbc.M312677200.

[89] U. Derewenda *et al.*, "The Crystal Structure of RhoA in Complex with the DH/PH Fragment of PDZRhoGEF, an Activator of the Ca2+ Sensitization Pathway in Smooth Muscle," *Structure*, vol. 12, no. 11, pp. 1955–1965, Nov. 2004, doi: 10.1016/j.str.2004.09.003.

[90] S. Xiang *et al.*, "The Crystal Structure of Cdc42 in Complex with Collybistin II, a Gephyrin-interacting Guanine Nucleotide Exchange Factor," *J. Mol. Biol.*, vol. 359, no.

1, pp. 35–46, May 2006, doi: 10.1016/j.jmb.2006.03.019.

[91] N. Mitin, L. Betts, M. E. Yohe, C. J. Der, J. Sondek, and K. L. Rossman, “Release of autoinhibition of ASEF by APC leads to CDC42 activation and tumor suppression,” *Nat. Struct. Mol. Biol.*, vol. 14, no. 9, pp. 814–823, Sep. 2007, doi: 10.1038/nsmb1290.

[92] S. Lutz *et al.*, “Structure of G q-p63RhoGEF-RhoA Complex Reveals a Pathway for the Activation of RhoA by GPCRs,” *Science (80-.).*, vol. 318, no. 5858, pp. 1923–1927, Dec. 2007, doi: 10.1126/science.1147554.

[93] K. Murayama, M. Kato-Murayama, R. Akasaka, T. Terada, S. Yokoyama, and M. Shirouzu, “Structure of the Rho-specific guanine nucleotide-exchange factor Xpln,” *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.*, vol. 68, no. 12, pp. 1455–1459, Dec. 2012, doi: 10.1107/S1744309112045265.

[94] Y. Shen *et al.*, “Crystal structure of the DH and PH-1 domains of human FGD5.”

[95] Z. Chen, L. Guo, S. R. Sprang, and P. C. Sternweis, “Modulation of a GEF switch: Autoinhibition of the intrinsic guanine nucleotide exchange activity of p115-RhoGEF,” *Protein Sci.*, vol. 20, no. 1, pp. 107–117, Jan. 2011, doi: 10.1002/pro.542.

[96] K. R. Abdul Azeez, S. Knapp, J. M. P. Fernandes, E. Klussmann, and J. M. Elkins, “The crystal structure of the RhoA–AKAP-Lbc DH–PH domain complex,” *Biochem. J.*, vol. 464, no. 2, pp. 231–239, Dec. 2014, doi: 10.1042/BJ20140606.

[97] A.-P. Petit, C. Garcia-Petit, J. A. Bueren-Calabuig, L. M. Vuillard, G. Ferry, and J. A. Boutin, “A structural study of the complex between neuroepithelial cell transforming gene 1 (Net1) and RhoA reveals a potential anticancer drug hot spot,” *J. Biol. Chem.*, vol. 293, no. 23, pp. 9064–9077, Jun. 2018, doi: 10.1074/jbc.RA117.001123.

[98] S. J. Bandekar *et al.*, “Structure of the C-terminal guanine nucleotide exchange factor module of Trio in an autoinhibited conformation reveals its oncogenic potential,” *Sci. Signal.*, vol. 12, no. 569, p. eaav2449, Feb. 2019, doi: 10.1126/scisignal.aav2449.

[99] C.-S. Goh, D. Milburn, and M. Gerstein, “Conformational changes associated with protein–protein interactions,” *Curr. Opin. Struct. Biol.*, vol. 14, no. 1, pp. 104–109, Feb. 2004, doi: 10.1016/j.sbi.2004.01.005.