UNIVERSITÀ DEGLI STUDI DI MILANO – BICOCCA

Dottorato di Ricerca in Scienze Chimiche – XXIII Ciclo

# COMPARISON OF PROTEIN DYNAMICS:

# A NEW METHODOLOGY BASED ON SELF-ORGANIZING MAPS

Tesi di Dottorato di:

Domenico Fraccalvieri

Tutor:

Prof. Laura Bonati

Coordinatore:

Prof. Franca Morazzoni

anno accademico 2009 – 2010

*Ai sognatori*

*To the dreamers*

# *C o n t e n t s*

# List of Figures

# List of Tables

# *Chapter 1*

## *INTRODUCTION*

Proteins are not static, they are flexible and sample a large ensemble of conformations around the average structure. Therefore a complete description of proteins requires, in addition to sequence and structure information, the knowledge of the multidimensional energy landscape that defines the different conformational states, their relative probabilities and the energy barriers between them. The most recent view of protein dynamics indicates that the dynamic landscape is an intrinsic property (or 'personality') of a protein. It is encoded in its fold and a ligand that interacts with the protein does not induce the formation of a new conformation but, instead, leads to a redistribution of the relative populations of conformational substates that already pre-exist in solution. This theory is called "conformational selection model" (1; 2)*.

The deep relation between dynamics, in terms of both global and local flexibility, and function of proteins is now widely acknowledged (1; 3-6). Flexibility is involved in protein binding to small molecules where, in many cases, only a conformational change in the binding site permits accommodation of the ligand. Large conformational changes, usually domain motions, are commonly observed in enzymatic catalysis and they are generally coupled to the interchange between the enzyme active and inactive forms, which can be triggered *e.g.* by substrate binding or phosphorylation. Protein dynamics is also at the basis of signal transduction processes and allosteric interactions. Moreover, there is growing evidence that intrinsic mobility is important in regulating protein-protein interactions. The hypothesis of the "conformational selection model", i.e. that proteins that exhibit substantial conformational change upon complexation have to

* = see also the references therein

be intrinsically flexible, has been confirmed, and in many studies the unbound protein was found to sample conformations close to the bound form (7). A fascinating feature of many proteins is multispecificity, i.e. the ability to bind several partners with good affinity. This implies that each interacting protein chooses a favourable counterpart as binding partner from the conformational ensemble of the receptor. In this framework, particular interest has raised on the dynamical properties of hub proteins and their transient complexes (8).

Moreover, it was recently suggested that intrinsic protein dynamics also defines the ability of proteins to adapt and evolve new functions (9). This idea motivated many studies in the last years, aimed at studying conservation and specialization of dynamic properties in protein families and superfamilies (10 -16). Another consequent implication is the possibility of using similarity between protein flexibilities to detect distant homologues (11).

In terms of timescale, it is possible to divide the dynamic processes into two categories: "fast" motions and "slow" motions. The "fast" motions include ps–ns dynamics of side chains and ns–μs loop and local hinge motions . The "slow" motions include μs–ms larger motions like collective domain motions and allosteric transitions. Furthermore, the shorter-timescale dynamics can influence and be influenced by longer timescale motions (1).

Dynamics on a "slow" timescale defines fluctuations between kinetically distinct states that are separated by energy barriers of several $kT$ (the product of the Boltzmann constant and the absolute temperature). These larger-amplitude collective motions between relatively small numbers of states are involved in many biological processes, including enzyme catalysis, signal transduction and protein–protein interactions. Nowadays both experimental and computational techniques can detect, and be used to describe, motions at these timescales. X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, cryo-electron microscopy and small-angle X-ray scattering provide atomic-resolution, or near-

atomic-resolution, snapshots of substates. At these timescales, the clear advantage of NMR methods is that they deliver the timescale of transitions, together with atomic resolution. The classical biophysical techniques, like fluorescence, circular dichroism, UV, IR and Raman spectroscopy, may provide kinetic information that is complementary to higher-resolution methods. Unfortunately, protein dynamics on the microsecond-to-millisecond timescale is currently out of reach for conventional Molecular-Dynamics (MD) simulations. To overcome this restriction, a large variety of approaches, like normal mode analysis and gaussian network models, have been developed to simplify force fields (17)*.

In the "fast" timescale dynamics, a large ensemble of structurally similar states that are separated by energy barriers of less than 1 $kT$ result in more-local, small-amplitude ps- to ns- fluctuations at physiological temperature. Also at this timescale experimental and computational techniques can be used to describe protein dynamics. An example of these descriptors are the B factors from X-ray diffraction, that may be associated to the mean-square atomic displacement, but it has to be considered that both true intramolecular motions and lattice disorder contribute to them. In the NMR relaxation methods, ps-to-ns dynamics are characterized in terms of the amplitude and the timescale of bond fluctuations. In solid-state NMR spectroscopy, motions on a broader timescale (low microsecond and faster) can be detected. At this timescale computational methods, and in particular Molecular-Dynamics simulations, are widely applied in the investigation of a wide range of dynamic properties and processes. Particularly in the last years, with advances in computational power, improved algorithms and reduced costs, MD simulations have become potent tools to investigate protein dynamics.

By means of MD simulations a large ensemble of molecular structures can be generated to sample the accessible conformational space of a protein. This allows

* = see also the references therein

the production and recording of time related configurations of the system as frames of a classical trajectory (18). After a stage of equilibration in the solvent, the system reaches an equilibrium state where, if a representative sampling of the phase space has been performed (19), the sampled structures describe a statistical ensemble that can be employed to derive average properties.

Recently, to help in obtaining new insights into protein function, simulations have been applied on a large scale. For example, in the context of a project called 'Dynameomics' (20), a database of c.a. 11000 molecular dynamics simulations of the native states and high-temperature unfolding pathways for over 2000* proteins has been built up. Another available database is MoDEL (Molecular Dynamics Extended Library) (21) in which more than 1700* trajectories of monomeric soluble structures are stored.

Drawbacks in the use of MD are that, on one side, some properties of interest are computationally time-consuming to evaluate and, on the other side, recurrence of transitions between conformations is difficult to extract from the raw ensemble data (22). For these reasons grouping the conformations becomes a necessity. The most desirable strategy would be to use *kinetic clustering* (23 - 28), where conformations are grouped according to their transition probabilities during the simulation and the identified clusters are directly related to the free energy landscape. A limitation to this approach arises from the need of an exhaustive sampling with convergence of all pairwise transition probabilities (28). A more affordable solution is to use *geometrical clustering*, because only a representative sampling of the accessible conformations is required. The underlying assumption is that structurally similar conformations lie in the same basin of the free energy surface. While often this is an acceptable approximation, a recent study suggested caution in interpreting the clustering results (28).

* = data updated on November 2010

*Geometrical clustering* for conformational analysis was introduced when simulation time increased up to nanoseconds generating tens of thousands of structures (29 - 31) and has been extensively used since then (22; 28). Several data-mining algorithms have been adopted but, according to a recent survey (22), no general strategy is available: clustering results are often influenced by the type of algorithm and the choice of optimal parameters is mostly left to the user experience and the specific case.

Originally adapted to analyze protein folding simulations (29) these algorithms were mainly implemented for multiple trajectories of the same system. However, in the last years a great interest has emerged in the comparison of protein flexibility of different proteins, with the main focus on comparing functionally related proteins or studying the evolutionary conservation and specialization of protein dynamics across distant homologous proteins (10 - 16). This new interest emphasises the need of more advanced tools to compare conformational ensembles of different protein domains especially when derived from extensive MD simulations.

For these purposes, data mining techniques, in particular neural network approaches can be interesting candidates. Nowadays computer technology has simplified the complexity in analyzing scientific data; visualizing data as colour-coded images that undergo qualitative changes to convey information for better pattern recognition many times provides a better understanding of the results, especially in studies for inferring inter-relationships (32). Within these techniques, the Self-Organizing Maps (SOMs or Kohonen maps) are an invaluable data mining tool (33). The SOM algorithm belongs to the unsupervised learning processes and it is based on similarity comparisons in a continuous space, which results in a system that associates similar inputs close to each other in the two-dimensional grid called “the map” . In principle, even if no explicit clusters exist in the data set, the output map reveals “ridges” and “ravines”. The former are open zones with

irregular shapes and high cluster tendency, whereas the latter separate data sets that have a different statistical nature. If data are close together in a "ridge" and they are connected, then they are similar. If they are not connected and they are separated by a "ravine", then they are different. This mapping complements the information on the grouping and opens possibility of further tool development.

SOMs were recently applied to conformational analysis of bio-molecules. A first application concerned the analysis of lipid molecules, where they resulted very effective in easily highlighting structural features, and distinguishing the main transitions from the minor conformational changes (34; 35). SOMs were also used to automatic clustering of protein-ligand docking poses (36).

While still not widely used in the analysis of conformational ensembles, SOMs are more accurate and provide more consistent results than traditional clustering algorithms (22). Moreover, they allow a topological mapping of the conformational space embedded in a simple 2D visualisation.

Aim of this thesis was the development of a novel and general approach to analyse and compare conformational ensembles of different protein domains using SOMs. The novelty of the approach concerns the application of this particular neural network algorithm to analyze data obtained by MD simulations. Once defined how to extract the data from MD trajectories and how to analyze them with a SOM, our effort was the definition of a general "rule" for the use the SOMs in this field.

First, we encoded the representation of the conformations extracted from the MD simulations as a proper input data for the SOM analysis.

Second, we studied the effect of the typical SOM learning and topological parameters in the analysis of these data, to define how to obtain reproducible and stable maps, by using an experimental design approach. This is a widely used technique to find out the best combination of design parameters and to reduce the variation for quality (37).

Third, concerning the non-solved problem of clustering a SOM, we proposed the use of a rule to define the optimal number of clusters that best summarizes the information in the map.

Finally, to evaluate the performance of the method in different comparisons of protein flexibility, we applied this protocol for the conformational and functional analysis of two study cases:

A group of single-site mutants of the α-spectrin SH3 (Spc-SH3) domain. This is an interesting case of a small intra-cellular signaling domain where ligand binding activity is modulated by single-mutations that greatly affect the conformational dynamics (38).

The bound and unbound states of a group of protein-protein complexes involving proteins of the RAS superfamily. These systems allowed to study transient complexes which show large conformational changes at the interface upon binding, along with the promiscuity of binding characteristic of hub proteins (8).

# *Chapter 2*

# METHODS

## *2.1 Molecular Dynamics simulations and analysis*

### *2.1.1 Molecular Dynamics*

Molecular Dynamics allows the production and recording of time related configurations of the system as frames of a classical trajectory. After an equilibration time the system reaches an equilibrium state and the sampled structures describe a statistical ensemble that can be employed to derive average properties. This condition is strictly related with the presence of a representative sampling of the phase space and it is expressed by the condition of ergodicity. Unfortunately, assessing extension and convergence of sampling for biomolecular simulations is not a trivial problem; it requires careful investigation and often an increase in computational efforts near to the limits of small laboratory resources (19).

In Molecular Dynamics the system is described as a collection of classical particles interacting via potentials with mainly pairwise components. Sets of coordinates are generated in a time sequence fashion through integration of Newton's equation of motion. The equation E1 presents the law of motion for a system of $N$ particles:

$$m_i \frac{\partial^2 \boldsymbol{r}_i}{\partial t^2} = \boldsymbol{F}_i \qquad i = 1, \dots, N \tag{E1}$$

where $\boldsymbol{r}_i$ are the positions and $\boldsymbol{F}_i$ the forces acting on them.

Forces acting on the system particles are the negative derivatives of the potential function of $\boldsymbol{r}$ (E2). This commonly takes an additive form composed by bonded and non-bonded terms (E3). As an example, the analytical form of the single terms of the GROMACS force field (39) are reported in equations E4-E8.

Electrostatic interactions can be calculated through Coulomb's law (E4): $q_i$ and $q_j$ are point charges on atom *i* and *j*, while $r_{ij}$ is their distances and $\epsilon_0$ and $\epsilon_r$ are in vacuum and relative dielectric constants. Van der Waals contributions can be described by a 12-6 Lennard-Jones potential (E5) where $A_{ij}$ and $B_{ij}$ are parameters depending on the pair of atoms *i* and *j*.

Bond and angle potentials (E6, E7) can take an harmonic form where $b_{ij}$ and $\theta^0_{ijk}$ are reference values and $k^b_{ij}$ and $k^\theta_{ijk}$ are force constants relative to the atom pairs or triplets.

Torsional interactions are calculated through a periodic function (E8) with a reference value $(\varphi^0_{ijkl})$, a parameter $(k^\varphi_{ijkl})$ that gives a qualitative indication of the relative barriers to rotation, and a multiplicity $(n_{ijkl})$ to include the number of minima in the function.

$$\boldsymbol{F}_i = -\frac{\partial V}{\partial r_i} \qquad V = V(\boldsymbol{r}_1, \dots, \boldsymbol{r}_n) \tag{E2}$$

$$V = V_{Coulomb} + V_{vdW} + V_{bonds} + V_{angles} + V_{dihedrals} \tag{E3}$$

$$V_{Coulomb} = \sum \frac{1}{4\pi\epsilon_0\epsilon_r}\frac{q_i q_j}{r_{ij}} \tag{E4}$$

$$V_{vdW} = \sum \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \tag{E5}$$

$$V_{bonds} = \sum_{bonds} \frac{1}{2} k^b_{ij}\left(r_{ij} - b_{ij}\right)^2 \tag{E6}$$

$$V_{angles} = \sum_{angles} \frac{1}{2} k^\theta_{ijk}\left(\theta_{ijk} - \theta^0_{ijk}\right)^2 \tag{E7}$$

$$V_{dihedrals} = \sum_{dihedrals} k^\varphi_{ijkl}\left(1 + \cos\left(n_{ijkl}\varphi_{ijkl} - \varphi^0_{ijkl}\right)\right) \tag{E8}$$

The potential form and the collection of parameters for each equation are usually referred as Force Field (FF). Different types of Force Field are available, each designed for specific aims and with a specific scope of application. Biomolecular simulations are performed with Force Fields that include optimised parameters for

amino acids, nucleic acid bases and small molecules of biological interest.
In the majority of cases potential terms for non-bonded interactions are described by pair-potentials while bonded interactions involve also three and four body terms. All interactions are short-ranged except electrostatic ones: these are computationally expensive and are usually approximated in the long range in different ways, ranging from brutal cut-off methods to more sophisticated procedure like Particle Mesh Ewald method (40).
Different versions of the GROMACS program (39; 41-44) were used for the simulations in this work (see Chapters 3 and 4 for the specific choices).
GROMACS supports a number of different Force Fields (43), including the GROMOS96 (45; 46), with different kind of parameterization (43a1, 43a2, 43b1, 43a3, 53a5, 53a6). GROMOS96 is a further development of the GROMOS87 FF, on which the GROMACS FF is based. It is a united atom FF, i.e. without explicit non-polar hydrogens.
The choice for this work was to employ the GROMOS96 43a2 FF (47), that is widely employed for protein simulations in water and is an improvement of the 43a1 (48), with a better description of alkanes.

### *2.1.2 System setup*

Protein structures are described in GROMACS with two kind of properties: static properties and dynamical properties.
The static properties are the Force Field parameters associated to atom types through a topological description of the system on a residue base. These properties are recorded in the topology file and remain unchanged for both simulation and analysis steps.
Positions and velocities of atoms are function of time. These dynamical properties are recorded in trajectory files and are the main output of the simulation process.

Starting positions are in a separated coordinates file. Topology file, coordinates file and a simulation parameter file are pre-processed in a binary file that constitutes the starting point for simulation.

A general protocol for system preparation was applied to all the proteins studied. The details of the protocol that were specifically set for each system will be described in Par 3.2.1, for the SH3 system, and in Par 4.2.1, for the RAS domains.
The main steps are:

1. topology and coordinates generation from the *.pdb* structure;
2. water box generation;
3. solvent equilibration;
4. ions addition;
5. short minimization.

The system topology and the description of the force field are read at the beginning and they are never modified during the simulation.
A cubic or octahedral box was generated around each protein, setting a minimum distance between solvent and box edges (see Fig. 2.1). This allowed the use of Periodic Boundary Conditions and minimum image convention during the simulations, with explicit description of solvent molecules. This is the classical way to minimize edge effects in a finite system. The atoms of the system to be simulated are put into a space-filling box, which is surrounded by translated copies of itself (Fig. 2.2). Thus there are no boundaries of the system; the artefact caused by unwanted boundaries in an isolated cluster is replaced by the artefact of periodic conditions. A simple point-charge (SPC) model (49) was used to describe the water molecules.

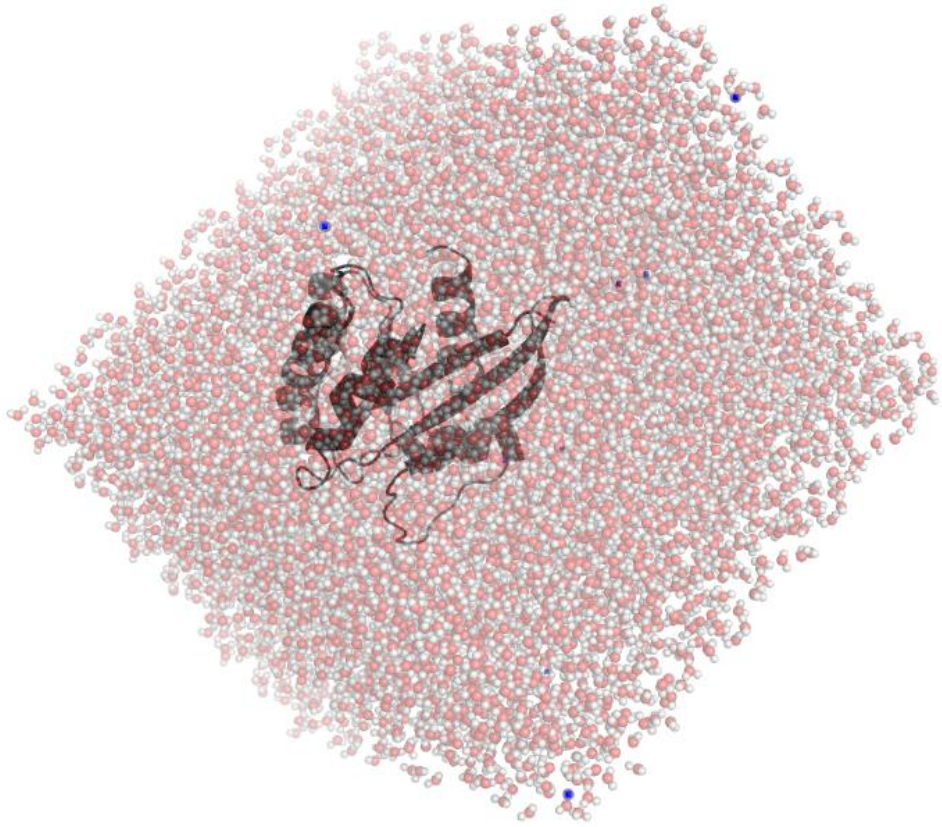


*Figure 2.1 – Example of protein immersed in a water cubic box. The structure of the protein is in black cartoons, the white and red spheres represent the water molecules, and the blue spheres the $Na^+$ ions added in the box to balance the overall charge of the system.*

The GROMACS algorithm replicates the box and deletes water molecules that can clash with the protein. This usually leads to a bad-description of water box that requires a further step of relaxation. This was reached through a short Molecular Dynamics simulation of few ps with positional restraints on the proteins. It is usually enough to remove the last clashes. During the molecular simulations the internal degrees of freedom of the solvent molecules are kept constant with the SETTLE algorithm (50).

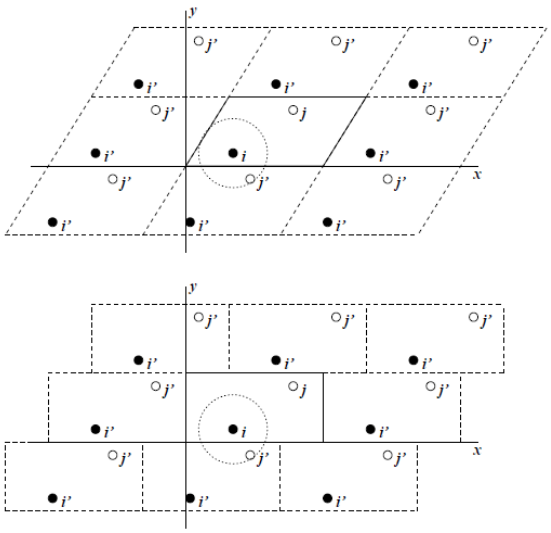



*Figure 2.2 – Periodic boundary conditions in two dimensions.*

Because often the structures of the proteins to simulate are negatively or positively charged, the addition of cations (e.g. $Na^+$) or anions (e.g. $Cl^-$) as counter-ions was required. This is necessary to avoid bad behaviour of charged group during neighbour list updates, like generation of fictitious dipoles.

The last step to prepare the system for simulation was a short Molecular Mechanics minimisation. Treatment of long-range interactions during minimisation was chosen to be consistent with that used in the next simulation step, i.e. the Particle Mesh Ewald method (40).

### *2.1.3 Molecular simulations*

Simulations and analysis of the data were conducted with reference to the starting point obtained through the setup steps described in the previous Paragraph. Each protein analysis was driven by comparison to this structure. This was aimed to gain a consistent start for all the proteins without regard to any lacks in the crystal structure. Protein motion timescales range from *ns* to *s* and most of the functional motions are visible only on order of *ms*. This leads to a obvious restriction in the ability to investigate these systems, because actual computational power allows

order of simulations up to 100s of ns for average proteins (100-200 amino acids). Everyday laboratory facilities usually restrict this to 10s of ns.

In the aims of this thesis there is the comparison of the dynamic properties of different domains, so each simulation was designed to achieve an efficient sampling of the energy landscape for the neighbourhood of starting structure (see Par. 3.2.1 and Par 4.2.1 for the details of the specific choices for each system).
Increase of experiment time required introduction of some computational methods to speed up the simulations. Regarding this, GROMACS allows a pretty stable way of reducing computational costs retaining effective and correct sampling.
Main time saving is gained through increase of timestep. This is dependent on ability of keeping an integration frequency accurate for monitoring force changes. This is easily achieved if the fastest degrees of freedom are uncoupled from the other motions and they do not have information that can be useful for the problem under investigation. In this case the really high frequency motions, as bond stretching, can be removed.
GROMACS way to achieve system simplification passes through use of constrained distances and removal of light particle motions. Main constrained distances methods are SHAKE (51) and LINCS (52). The latter is usually employed in GROMACS for bond constraints. It is a non-iterative two steps algorithm that resets bond length after an unconstrained step. It is usually more stable than SHAKE and account for increase in timestep up to 2fs.
Non-bonded interactions and especially electrostatic interactions are important in description of complex systems, such as biomolecules. GROMACS supports fast but still remarkably accurate algorithms for this class of interactions.
With Periodic Boundary Conditions a grid-based neighbour list search allows a dynamically update of pair-lists. For each particle two cutoff distances are defined, to identify which contributions must be calculated at each step and which are updated only every n (usually set to 10) steps with the neighbour list itself. Long-

range electrostatic interactions are supported with different level of approximation: single and double cutoffs, Reaction Field, Ewald summation and Particle Mesh Ewald method.

In all the simulations of this thesis the Particle Mesh Ewald (PME) method (40) was employed. It is based on the Ewald summation method developed originally for crystal systems. The original method allows to transform the slow convergent summation of electrostatic pairwise contributions over the whole space to three fast convergent terms: a constant term and two sums, one in the direct space and one in the reciprocal space. Particle Mesh Ewald improves the performance of the calculation in the reciprocal space. The results is the possibility to retain relatively short cutoffs with higher accuracy in the electrostatic contribution.

Initial atomic velocity for all proteins were generated from a Maxwellian distribution at T = 300K (E9), where $p(v_i)$ is the probability for velocity $v_i$, $m_i$ is the particle mass and $k$ is the Boltzmann's constant. The program indeed implement the generation of this set of velocities consistent with an absolute value of temperature.

$$p(v_i) = \sqrt{\frac{m_i}{2\pi kT}}\, e^{-\frac{m_i v_i^2}{2kT}} \qquad \text{(E9)}$$

During the simulation the centre of mass of the system was removed and the system was weakly coupled with a thermal bath. The temperature was coupled groupwise, both protein and solvent are reset to 300K every 0.1 ps. The choice of the Berendsen thermostat is motivated by a more correct description of kinetics and higher stability of temperature coupling. Another enhancement of stability for long timestep is achieved by the specific implementation of the leap-frog (53) Molecular Dynamics integrator in GROMACS (54).

Leap-frog needs positions $\boldsymbol{r}$ at time $t$ and velocities $v$ at time $t - \frac{\Delta t}{2}$. From the positions, forces $\boldsymbol{F(t)}$ are updated and used to compute $\boldsymbol{v}\left(t + \frac{\Delta t}{2}\right)$ and then $\boldsymbol{r}(t + \Delta t)$:

$$\boldsymbol{v}\left(t+\frac{\Delta t}{2}\right)=\boldsymbol{v}\left(t-\frac{\Delta t}{2}\right)+\frac{\boldsymbol{F}(t)}{m}\Delta t \quad \text{(E10)}$$

$$\boldsymbol{r}(t+\Delta t)=\boldsymbol{r}(\boldsymbol{t})+\boldsymbol{v}\left(t+\frac{\Delta t}{2}\right)\Delta t \quad \text{(E11)}$$

### *2.1.4 Sampling Analysis*

Sampling analysis was aimed to identify the extension of sampling in the simulation. The efficiency of computer simulation in sampling the minima of proteins is still a problem, particularly evident if the analysis is directed to discover collective motions that can be related to biological functions. In fact, previous reports demonstrated that insufficient sampling can lead to mistake patterns of random diffusion for functional motions (19). To address this problem, different approaches can be adopted.

The most employed tool to evaluate the sampling is the Root Mean Square Deviation (RMSD) of a set of atoms with respect to a reference structure. The procedure requires a least-square fitting to the reference structure and then the calculation of RMSD:

$$RMSD_k(t_1,t_2)=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\|r_i(t_1)-r_i(t_2)\|^2} \quad \text{(E18)}$$

For a system of *N* atoms, $r_i(t_1)$ is the position of atom *i* at time $t_1$. For proteins the fitting is usually done only on the backbone (N,Cα,C) and the RMSD values are computed on all the protein atoms. If the reference is the starting structure ($t_2 = 0$), the information can give an idea of the time of equilibration and the presence of big conformational changes. Anyway the simple RMSD is always a measure strongly affected by the presence of small set of atoms with high differences from the reference structure and the fitting is not a problem with exact solution.

A more informative tool is the RMSD matrix. This reports the RMSD value for all the pairs of frames in the trajectory. From this picture one can highlight the re-sampling of the same substructure and the presence of transitions.

The overlap between the conformational space spanned by different parts of the simulation is also frequently used as an index for sampling convergence (19). This can be extracted directly from the information contained in the covariance matrix of the atomic coordinates (see Par. 2.1.5). In general, the overlap between two matrices **A** and B, $s(\mathbf{A},\mathbf{B})$, can be defined as:

$$s(\boldsymbol{A},\boldsymbol{B}) = 1 - \frac{\sqrt{tr((\boldsymbol{A}^{1/2} - \boldsymbol{B}^{1/2})^2)}}{\sqrt{tr\boldsymbol{A} + tr\boldsymbol{B}}} \qquad \text{(E12)}$$

where *tr* is the trace of the matrix. When $s(\mathbf{A},\mathbf{B})$ is equal to 1 the two spanned subspaces are identical, whereas a value of 0 indicates complete orthogonality. In this thesis, the overlap between the covariance matrix of each half of a simulation and the overall trajectory was evaluated for each protein studied.

### *2.1.5 Essential Dynamics*

Essential Dynamics (ED) is a widely applied technique based on principal component analysis (PCA) of conformational ensembles (55), aimed to extract informative directions of motion in a multidimensional space. It allows to both reduce the overall complexity of the simulation and isolate the important motions with a putative functional meaning.

The method is rooted on the basic assumption that the dynamics of a protein structure can be decomposed in this two contributions: an essential informative motion with possible biological interest and a constrained useless noise-like vibration. The Essential Dynamics, or Covariance analysis, requires the following steps:

1. Covariance matrix is constructed to describe atomic motions correlation. The element $C_{ij}$ for the *i* and *j* couple of coordinates is:

$$C_{ij} = \left(\overline{x_i - \overline{x_i}}\right)\left(\overline{x_j - \overline{x_j}}\right) \quad \text{(E12)}$$

Over-lines denote averages over all the data. In the case of a protein simulation,

those can be frames of a trajectory while the coordinates are the 3N coordinates of the system or a subset of them;

2. The covariance matrix can be diagonalised with an orthonormal transformation matrix ***R***:

$$R^T C R = diag(\lambda_1, \lambda_2, \dots, \lambda_N) \quad \text{(E13)}$$

The column of the R rotational matrix are the eigenvectors defining principal modes. The relative eigenvalues $\{\lambda_i\}$ are the variance of the data set in the new directions. They are usually sorted in decreasing order. For molecular simulations the eigenvalues contain the amount of fluctuations described by each new spatial direction. Indeed the sum of the eigenvalues describes the total positional fluctuation of atoms included in the covariance analysis:

$$\sum_i \left(\overline{x_i - \overline{x_i}}\right) = \sum_i (\lambda_i) \quad \text{(E14)}$$

Projection of original data on the eigenvectors generates the principal component:

$$\boldsymbol{p} = R^T(x - \overline{x}) \quad \text{(E15)}$$

Where p are the 3N principal components for the set of data x;

3. Evaluation of the amount of variance contained in the first eigenvectors leads to a separation of the new ***p*** space in two subspaces:

Essential Subspace: $\{q_i, \dots, q_n\}$ (E16)

Constrained Subspace: $\{q_{n+1}, \dots, q_{3N}\}$ (E17)

4. The data set can be described in a lower dimensional space. For protein simulations the dynamics can be analyzed in a subspace that is usually smaller than 5% of the original space.

It was demonstrated that the analysis can be performed only on the Cα atoms, because in most of the cases this reduction can retain the relevant information about functional protein dynamics (55).

In order to separate Essential from Constrained Subspace, it is necessary to identify the number of eigenvectors necessary to describe the protein system. This can be assessed looking at two measures: the amount of motion included in the Essential Subspace and the distribution of motion along the directions of the Essential subspace. The sum of eigenvalues for the eigenvectors defining Essential Subspace is a good index of the amount of motion included in the subspace. If this is compared to the total displacement fluctuation in the system it is possible to assess also the extent of separation between the two subspaces. In Essential Dynamics applied to Molecular Dynamics trajectories the amount of eigenvectors needed to explain around 70-80% of motion is usually around 20-30. The distribution of motion along the eigenvectors tends to be anharmonic with two or more peaks on the first directions and with a shaped of narrow gaussian on the last ones. The amplitude is usually broad on first and becomes more and more narrow on the following directions. Defining a number of directions is important to extract a filtered description of MD simulation, but also to compare information from different simulations.

In this thesis, all the analysis of conformational flexibility of protein domains were performed, after ED analysis, in the essential subspace.

### *2.1.6 Analysis of conformational flexibility*

To analyze the local flexibility of each domain the root mean square fluctuation (RMSF) on the positions of the Cα atoms, as obtained from the coordinate of the system in the essential subspace, was calculated. The RMSF of the Cα atom *i* is a measure of the mean deviation between its positions during the trajectory ($\boldsymbol{r}_j(t_j)$) and its time-averaged position ($\overline{\boldsymbol{r}_j}$):

$$RMSF_i = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(\boldsymbol{r}_j(t_j) - \overline{\boldsymbol{r}_j}\right)^2} \qquad \text{(E19)}$$

where N is the number of conformation analyzed in the trajectory.
Only Cα atoms were included in the analysis.

When the interest is to analyse the fluctuation of specific region of the domain (for example, the binding site) other indices can be employed. In this thesis the geometry of the protein binding site was described by a selected set of pairwise atomic distances, and its conformational changes were measured by the distance root mean square deviation (dRMSD) between the average conformation in the MD trajectory, a, and a reference structure, b:

$$dRMSD_{ab} = \sqrt{\frac{\sum_i \sum_j \left(d_{ij}^a - d_{ij}^b\right)^2}{N}} \qquad \text{(E20)}$$

(d is the distance value, i and j the indices of the selected atoms, and N is the total number of distances).
This index has the advantage that, at difference to the standard RMSD, no fit is needed.

## *2.2 Self-Organizing Maps (SOMs)*

### *2.2.1. Unsupervised Competitive learning*

A self-organizing map (SOM) (33), also called self-organizing feature map (SOFM) or Kohonen map, is a type of artificial neural network that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized representation of the multidimensional input space of the training samples. This bidimensional representation is called map. In the unsupervised learning there is no teacher; i.e. no feedback is allowed from the environment to say what the output should be or whether it is correct. The network must discover for itself patterns, features, regularities, correlations or categories in the input data and code them in the output space. The patterns which can be detected through an unsupervised learning network depend on the architecture; there are a number of possibilities, such as: familiarity, PCA (principal components analysis), clustering, prototyping, encoding, feature mapping. This is not an exhaustive classification of network architectures, basic architectures can be combined in several ways. In principle the unsupervised learning architectures are fairly simple, the complications and subtleties come mainly from the learning rules.
In particular, SOMs are unsupervised competitive learning methods; all output units compete for being the one to fire, only one fires and it is usually called the winner-take-all unit. The general aim of this network is to cluster the input data. The rationale is that similar input should be classified as being in the same category, and so should fire the same output unit.

A closely related topic is feature mapping, that is distinguished by the development of significant spatial organization in the output layer. In these networks the location of the winning output convey some information, nearby output correspond to nearby input patterns. More technically, if $\boldsymbol{\xi}^1$ and $\boldsymbol{\xi}^2$ are two input vectors, and $\mathbf{r}^1$ and $\mathbf{r}^2$ are the locations of the corresponding winning output

neurons, then $\mathbf{r}^1$ and $\mathbf{r}^2$ should get closer and closer, eventually coinciding, as $\boldsymbol{\xi}^1$ and $\boldsymbol{\xi}^2$ are made more similar and similar. What we are asking for is a topology preserving map from the space of possible inputs to the line, or plane, of the output units. A topology preserving map is essentially a map that preserve the neighborhood relations. A complete and exhaustive review of this topic can be found in tow books: *Introduction to the theory of neural network computation* (56) and in *Neural Network Design* (57).

### *2.2.2. The SOM algorithm*

A SOM consists of neurons organized on a regular low-dimensional, usually bidimensional, grid. The number of neurons may vary from few units up to thousands. Each neuron is represented by a d-dimensional weight vectors, $\underline{w} = [w_1, \dots, w_d]$, where d is the dimension of the input data vectors. The neurons are connected to adjacent neurons by a neighborhood relation (Fig. 2.3), which dictates the topology of the map (Fig. 2.4). The learning algorithm could be divided in five steps:

1) Randomize the initial weight vectors of the map's nodes
2) Grab an input data
3) Traverse each node in the map
4) Update the best match unit (BMU) and its neighborhoods by pulling them closer to the input vector
5) Repeat from 2 for each of the data for a number of epochs to reach a convergence of the map.

In this work, the practical applications of these step was performed using the SOM Toolbox 2.0 for Matlab (58). The details of each step are the following.

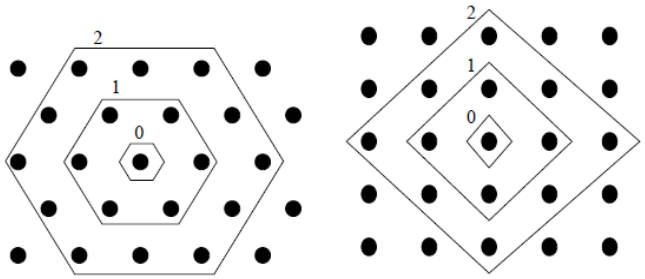



*Figure 2.3: SOM discrete neighborhood and lattice: the black dots represent the neurons of the map with hexagonal (on the left) and rectangular (on the right) lattice. The polygons between the neurons correspond to different neighborhood radius (from 0 to 2).*

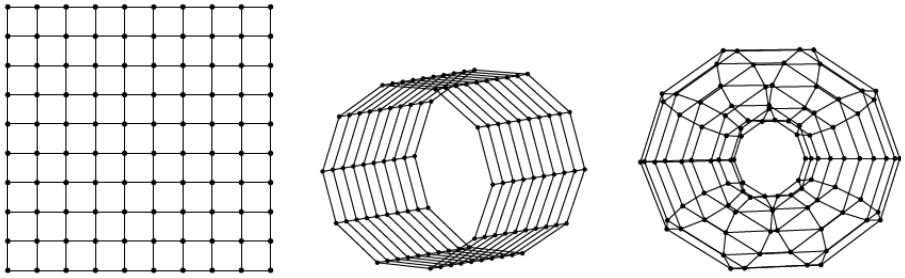


*Figure 2.4: SOM shapes: sheet (on the left), circular (in the middle) and toroid (on the right).*

- *Randomize the initial weight vectors of the map's nodes*

Before the training starts, initial values are given to the prototype vectors ($\underline{w}$). Typically one of the three following initialization procedures is used:

- ✓ random initialization, where the weight vectors are initialized with small random values;
- ✓ sample initialization, where the weight vectors are initialized with random samples drawn from the input data set;
- ✓ linear initialization, where the weight vectors are initialized in an orderly fashion along the linear subspace spanned by the two principal eigenvectors of the input data set.

- *Grab an input data*

Each input data vector is presented to the map. To avoid that the order in which the data are presented influence the resulting output map, in each epoch the data are randomly selected.

- *Traverse each node in the map*

In this step two sub-processes are involved.

First, a function $(d)$ is used to find the similarity/distance between the input vector $(\underline{x})$ and the map's node's weight vector $(\underline{w})$.

To evaluate the distance, different approaches can be used. In the SOM toolbox the Euclidean distance (E21) is given as default.

$$d(\underline{x}, \underline{w}) = \sqrt{(\sum_{i=1}^{n}(x_i - w_i)^2)} \quad \text{(E21)}$$

Second, the tracking of the node that produces the smallest distance (this node is the best matching unit, BMU).

Once evaluated the distances between the input data $(\underline{x_1})$ and each node of the map, the BMU is the neuron which satisfies the following criteria:

$$BMU_1 = \min(d(\underline{x}_1, \underline{w})) \quad \text{(E22)}$$

- *Update the BMU and its neighborhoods by pulling them closer to the input vector*

The neighbourhoods can be defined using four different functions (see Fig. 2.5); *bubble* (E23), *gaussian* (E24) , *cutgauss* (E25)and *ep* (E26).

To briefly describe each function, we have to define the neighbourhood radius at time $t$ $(\sigma_t)$, the distance between two unit of the map $a$ and $b$ $\left(d_{a,b}\right)$ and the step function $\left(1(x)\right)$.

The step function is equal to 0 if $x < 0$ and to 1 if $x \geq 0$.

Using these definitions, the functions are:

$$h_{ab}(t) = 1(\sigma_t - d_{ab}) \quad \text{(E23)}$$

$$h_{ab}(t) = e^{-d_{ab}^2/2\sigma_t^2} \quad \text{(E24)}$$

$$h_{ab}(t) = e^{-d_{ab}^2/2\sigma_t^2} 1(\sigma_t - d_{ab}) \quad \text{(E25)}$$

$$h_{ab}(t) = \max\{0, 1 - (\sigma_t - d_{ab})^2\} \quad \text{(E26)}$$

During the training phase, the learning rate $\left(\alpha(t)\right)$, i.e. the weight applied to the update process during the training, is a decreasing function which ranges between 1 and 0. A value of $\alpha_t$ close to 1 is used to implement fast learning, while more

conservative values of $\alpha_t$ are used in the final steps of the learning process. Also for this parameter multiple decreasing functions are available: *linear* (E27), *power* (E28), and *inv* (E29)

$$\alpha(t) = \alpha_0 \left(1 - \frac{t}{T}\right) \quad \text{(E27)}$$

$$\alpha(t) = \alpha_0 \left(\frac{0.005}{\alpha_0}\right)^{\frac{t}{T}} \quad \text{(E28)}$$

$$\alpha(t) = \frac{\alpha_0}{\left(1+100\frac{t}{T}\right)} \quad \text{(E29)}$$

Where $T$ is the length of the training, in terms of epochs, and $\alpha_0$is the initial learning rate.

The update process can be done using two algorithms: sequential and batch.

In the *sequential* algorithm after finding the BMU, the prototype vectors of the SOM are updated. The prototype vectors ($BMU_i$) and its topological neighbors are moved closer to the input vector in the input space (see Fig 2.6). The update process is described by the following formula (E30):

$$BMU_i(t+1) = BMU_i(t) + \alpha(t) h_{ab}(t) \left[\underline{x_i}(t) - BMU_i(t)\right] \quad \text{(E30)}$$

In the *batch* algorithm, the whole training set is gone through at once and only after this the map is updated with the net effect of all the samples (E31):

$$BMU_i(t+1) = \frac{\sum_{j=1}^{n} h_{ab}(t) \underline{\underline{x_j}}}{\sum_{j=1}^{n} h_{ab}(t)} \quad \text{(E31)}$$

All the steps above described will be repeated for a given number of epochs ($T$) sufficient to ensure a convergent training of the map (33).

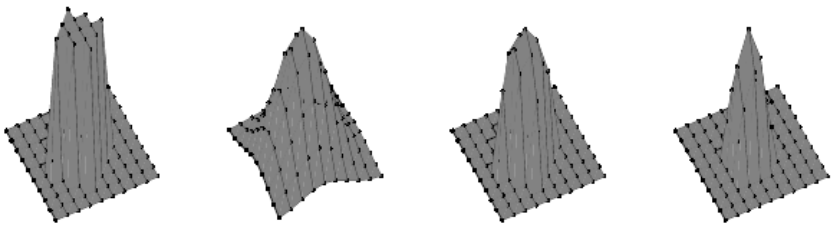


*Figure 2.5: SOM neighborhood functions: different neighborhood functions on a 2D map grid. From the left "bubble", "gaussian", "cutgauss" and "ep" ($\sigma_t$ = 2 is used as example).*

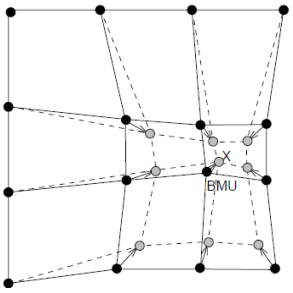



*Figure 2.6: SOM updating process of the BMU: the updating process of the BMU, in the sequential algorithm, and its neighborhoods after the learning of the input vector X. The solid line represent the starting situation, the dashed line the situation after the updating. The lines between the neurons (black dots) represent the distance between them.*

## *2.3 Development of a protocol to optimize the SOMs for structural ensemble analysis*

The major aim of this thesis is the development of a novel approach for conformational analysis of structural ensembles using SOMs. In particular, a protocol was designed to compare MD trajectories of protein domains (59).
The protocol consists of three steps. First, the SOM's parameters optimization is performed by experimental design. Then, the optimal sampling rate of the MD trajectory is calculated. Finally, the representative conformations projected onto the SOM's output are clustered. A diagram of the protocol is shown in Fig. 2.7.
In the next paragraphs the detail of the protocol is discussed, i.e. how to derive SOM input vectors from ensembles of conformations extracted from MD trajectories of one or more domains; how to find the optimal SOM, i.e. the SOM model which "best characterizes" the underlying input space structure; how to define the minimum number of structures to be selected and given as input data to the SOM, while maintaining a reliable picture of the protein dynamics; how to automatically define the number of clusters that best summarize the map obtained.

Numerical experiments concerning a study case composed by the α-spectrin SH3 (Spc-SH3) protein domain and a group of its single mutants will be presented in Par 2.3.7. The dynamical properties of these systems and their role in the SH3 functionality will be presented and discussed in Chapter 3.

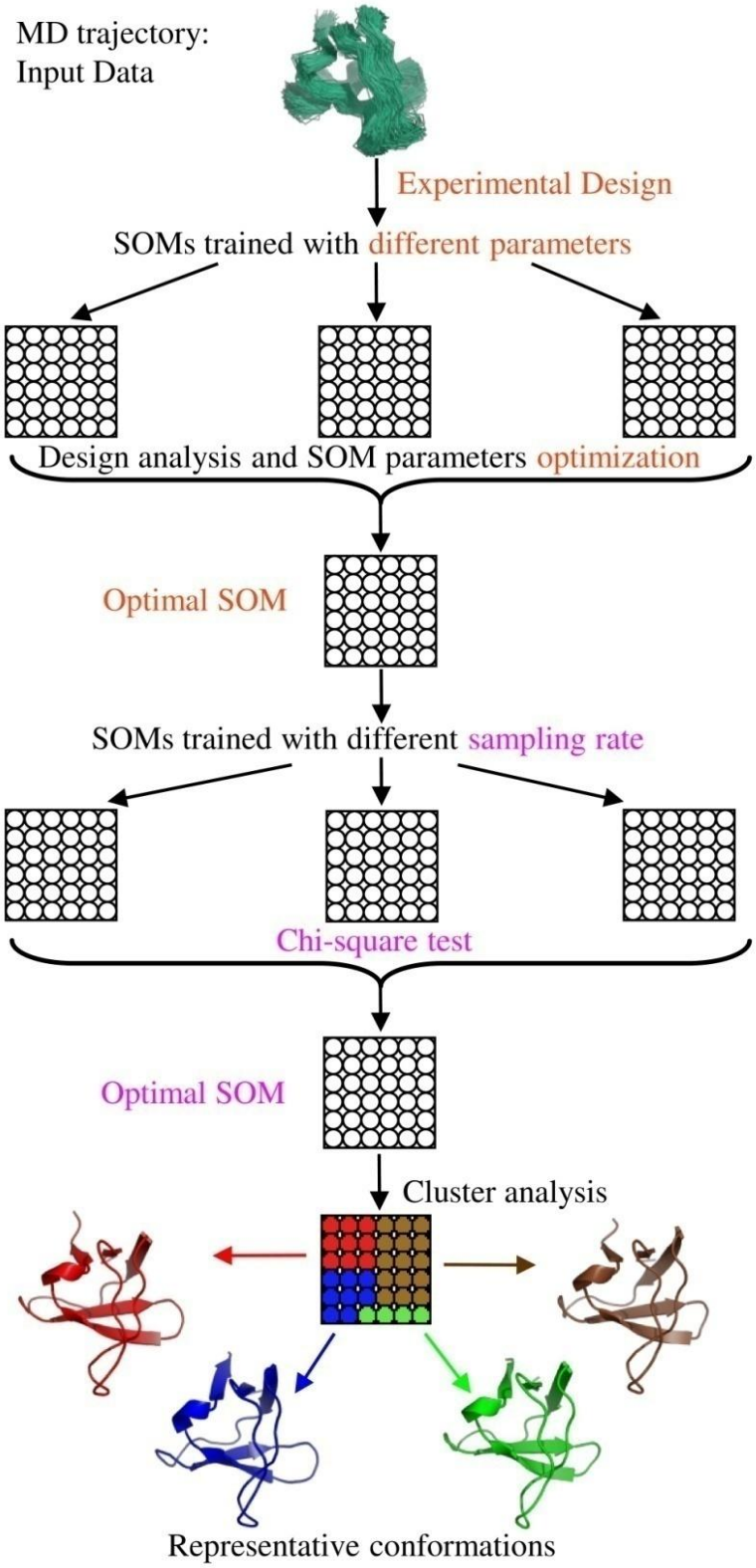



*Figure 2.7 - Optimization of a Self-Organising Map for structural ensemble analysis: diagram of the proposed protocol to optimise SOMs for structural ensemble analysis.*

### *2.3.1 Input data: from MD trajectory to SOM input vectors*

As described in Par. 2.2, the SOM is a neural network able to represent high-dimensional data onto a low dimensional lattice in a topology-preserving manner. Instead, as described in Par 2.1, Molecular Dynamics allows the production and recording of time-related configurations of a molecular system as frames of a classical trajectory.

Therefore, it should be clear that the first problem to deal with is to create a representation of the conformations produced with an MD simulation that is suitable to be learned by a SOM.

The steps proposed to this aim are:

1. To produce a MD trajectory of a given system for a given time length;
2. To evaluate the completeness of the sampling (see Par. 2.1.4);
3. To define the Essential Space (ES), useful to describe the most informative motions (see Par. 2.1.5 for details);
4. To extract the conformations from the trajectory, with a given sampling rate, and to record the Cartesian coordinates of the Cα, projected in the ES, in *.pdb* files;
5. To transform each *.pdb* file into a vector. The vector $(\underline{x}_{t=i})$, is associated with the i$^{th}$ frame of the trajectory $(t = 1, \dots, T)$, formed by the Cartesian coordinates of the *n* Cα of the system;
6. To create a matrix $\boldsymbol{X} = \begin{bmatrix} \underline{x}_{t=1} \\ \cdots \\ \underline{x}_{t=T} \end{bmatrix}$, where each vector $(\underline{x}_{t=i})$ is labeled using the progressive number of the i$^{th}$ frame;
7. To input $\boldsymbol{X}$ to the SOM; during each learning epoch, the vectors $(\underline{x}_{t=i})$ are randomly extracted from $\boldsymbol{X}$.

### *2.3.2 Design of experiments*

The result of the SOM learning process depends on several design parameters and the problem to find the optimal SOM, i.e. the SOM model which "best characterizes" the underlying clusters, is extremely complex. It is an ill posed problem (33) and several efforts have been devoted to provide an efficient and effective solution (60; 61). The problem consists of selecting the value of the SOM design parameters which bring to the optimal SOM. This problem is time consuming. Indeed, it is known that Neural Networks are affected by the Curse of Dimensionality (62). Analysts can reduce the impact of the Curse by several strategies, including design of experiments, extracting low-dimensional features, imposing parsimony, or aggressive variable search and selection.

Using the SOM toolbox 2.0 for Matlab (58), the following SOM design parameters are available:

- Topological parameters: map size, lattice type, shape;
- Parameters associated to the learning process: learning algorithm, neighborhood function, alpha type, radius, training length and starting alpha.

The ranges of the SOM parameters are summarized in Table 2.1.

To briefly describe each parameter:

- Map size: the number of neurons, usually disposed in a square grid;
- Lattice type: the local lattice structure. A rectangular neuron borders with four other neurons, an hexagonal one with six (see Fig. 2.3);
- Shape: the global shape of the map (see Fig. 2.4).
- Learning algorithm: the update process used to learn the input output mapping;
- Neighbor function: the shape of the neighborhood function on the map (see Fig. 2.5);

- Alpha type: the decreasing function of the alpha values, as described in Par 2.2.2;
- Radius: the neighborhoods' kernel, as described in Par 2.2.2;
- Training length: the number of epochs of the training process;
- Starting alpha: the value of $\alpha_0$, as defines in Par 2.2.2.

*Table 2.1: SOM design paramater: SOM's design parameters together with allowed ranges*

| Design parameter | Range |
|---|---|
| Map size | [100, 400] |
| Lattice type | {hexagonal, rectangular} |
| Shape | {sheet, cylinder, toroid} |
| Learning algorithm | {batch, sequential} |
| Neighbour function | {gaussian, bubble, ep} |
| Alpha type | {inverse, linear, power} |
| Radius | {1, 2, 3} |
| Training length | [1000, 5000] |
| Starting alpha | [0.01, 0.09] |

In this work, design of experiments (37) is used to model the unknown mapping between the design parameters and the performance function. Once such mapping has been built it is possible to find the optimal value of the SOM design parameters, i.e. to discover the optimal SOM. There are numerous design criteria that have been studied: D-Optimality, which minimizes the variance of the parameters estimates, G-Optimality, which minimizes the maximum prediction variance. However, in this case the particular nature of design parameters does not allow the implementation of an experimental design plan based on one of the above criteria.

Therefore the Taguchi robust design plan (63) has been implemented. Such design plan, arranges variables or factors in an orthogonal array. The orthogonal array properties are such that between each pair of columns each combination of levels

(or variables) appears an equal number of times. Due to an orthogonal layout, the effects of the other factors can be balanced and give a relative value representing the effects of a level compared with the other levels of a given factor. In orthogonal array experiments, the number of test runs is minimized, while keeping the pair-wise balancing property. The principle is that statistically planned experiments are essential for a successful parameter design. Taguchi methods utilize two-, three-, and mixed level fractional factorial designs, and are used for maximizing robustness of products and processes, thereby achieving high quality at a low cost (63).

The experimental design plan that was used in the work consists of 36 runs with 3 replicas each, thus an overall number of experiments equal to 108 for each training dataset. The values of design parameters included in the Taguchi design plan are summarized in Table 2.2.

*Table 2.2: Set of the 36 experiments defined using the Taguchi method. Each row of the table contains the specific values used in the experiment.*

| Map size | Lattice type | Shape | Learning algorithm | Neighbour function | Alpha type | Radius | Training length | Starting alpha |
|---|---|---|---|---|---|---|---|---|
| 100 | hexagonal | sheet | batch | Gaussian | inverse | 2 | 1000 | 0.01 |
| 225 | hexagonal | cylinder | batch | Bubble | power | 3 | 3000 | 0.05 |
| 400 | hexagonal | toroid | batch | Ep | linear | 1 | 5000 | 0.09 |
| 100 | hexagonal | sheet | sequential | Gaussian | linear | 3 | 1000 | 0.09 |
| 225 | hexagonal | cylinder | sequential | Bubble | inverse | 1 | 3000 | 0.01 |
| 400 | hexagonal | toroid | sequential | Ep | power | 2 | 5000 | 0.05 |
| 100 | rectangular | cylinder | sequential | Gaussian | linear | 1 | 5000 | 0.05 |
| 225 | rectangular | toroid | sequential | Bubble | inverse | 2 | 1000 | 0.09 |
| 400 | rectangular | sheet | sequential | Ep | power | 3 | 3000 | 0.01 |
| 100 | rectangular | toroid | batch | Gaussian | power | 1 | 3000 | 0.01 |
| 225 | rectangular | sheet | batch | Bubble | linear | 2 | 5000 | 0.05 |
| 400 | rectangular | cylinder | batch | Ep | inverse | 3 | 1000 | 0.09 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 100 | hexagonal | toroid | sequential | Bubble | power | 1 | 1000 | 0.05 |
| 225 | hexagonal | sheet | sequential | Ep | linear | 2 | 3000 | 0.09 |
| 400 | hexagonal | cylinder | sequential | Gaussian | inverse | 3 | 5000 | 0.01 |
| 100 | rectangular | toroid | sequential | Bubble | inverse | 3 | 3000 | 0.09 |
| 225 | rectangular | sheet | sequential | Ep | power | 1 | 5000 | 0.01 |
| 400 | rectangular | cylinder | sequential | Gaussian | linear | 2 | 1000 | 0.05 |
| 100 | rectangular | sheet | sequential | Bubble | linear | 3 | 5000 | 0.01 |
| 225 | rectangular | cylinder | sequential | Ep | inv | 1 | 1000 | 0.05 |
| 400 | rectangular | toroid | sequential | Gaussian | power | 2 | 3000 | 0.09 |
| 100 | rectangular | cylinder | batch | Bubble | power | 2 | 5000 | 0.09 |
| 225 | rectangular | toroid | batch | Ep | linear | 3 | 1000 | 0.01 |
| 400 | rectangular | sheet | batch | Gaussian | inverse | 1 | 3000 | 0.05 |
| 100 | hexagonal | cylinder | batch | Ep | power | 2 | 1000 | 0.01 |
| 225 | hexagonal | toroid | batch | Gaussian | linear | 3 | 3000 | 0.05 |
| 400 | hexagonal | sheet | batch | Bubble | inverse | 1 | 5000 | 0.09 |
| 100 | hexagonal | cylinder | batch | Ep | linear | 1 | 3000 | 0.09 |
| 225 | hexagonal | toroid | batch | Gaussian | inverse | 2 | 5000 | 0.01 |
| 400 | hexagonal | sheet | batch | Bubble | power | 3 | 1000 | 0.05 |
| 100 | hexagonal | toroid | sequential | Ep | inverse | 3 | 5000 | 0.05 |
| 225 | hexagonal | sheet | sequential | Gaussian | power | 1 | 1000 | 0.09 |
| 400 | hexagonal | cylinder | sequential | Bubble | linear | 2 | 3000 | 0.01 |
| 100 | rectangular | sheet | batch | Ep | inverse | 2 | 3000 | 0.05 |
| 225 | rectangular | cylinder | batch | Gaussian | power | 3 | 5000 | 0.09 |
| 400 | rectangular | toroid | batch | Bubble | linear | 1 | 1000 | 0.01 |

### *2.3.3 Objective function: the distance minimum.*

The parameters which are detected to be relevant by design of experiments are used to fit a linear regression model which links their values to a clustering performance measure (objective function), and thus to select the optimal values of the SOM design parameters. The selected objective function (64) is:

$$\frac{\sum_j \sum_{i \in C_j} d\left(\underline{w}_{ij}, \underline{\mu}_j\right)}{R_{tot}} + \sum_j d\left(\underline{\mu}_j, \underline{\mu}\right) \qquad \text{(E32)}$$

where j and $C_j$ are the j$^{th}$ cluster of neurons and the set of associated neurons; $\underline{w}_{ij}$ is the weight vector associated with the i$^{th}$ neuron of the j$^{th}$ cluster; $\underline{\mu}_j$ is the centroid of the j$^{th}$ cluster (the mean vector whose components are the arithmetic averages of the components of $\underline{w}_{ij}$); and $\underline{\mu}$ is the centroid of the overall map (the mean vector whose components are the arithmetic averages of all the weight vectors of the map). The distance (d) is Euclidean and $R_{tot}$ is a normalization factor equal to:

$$R_{tot} = \frac{\#\ of\ neurons\ of\ the\ experiment}{100}$$

As shown in Table 2.2, in the experiments here performed the number of neurons was set to 100, 225, and 400, so $R_{tot}$ is 1, 2.25, and 4.

This normalization factor affects only the first addend of the original equation (64), making possible the comparison of maps with different number of neurons. In fact, the first term in (E32), that is the sum of the distances of each neuron to the centroid of its cluster, increases with the number of neurons. The effect of the normalization factor is shown for a case study in Fig. 2.8, where the two terms in (E32) are calculated with and without $R_{tot}$. The second term of (E32), that compute the inter cluster distance (in the middle of Fig. 2.8), is not influenced by the number of neurons in

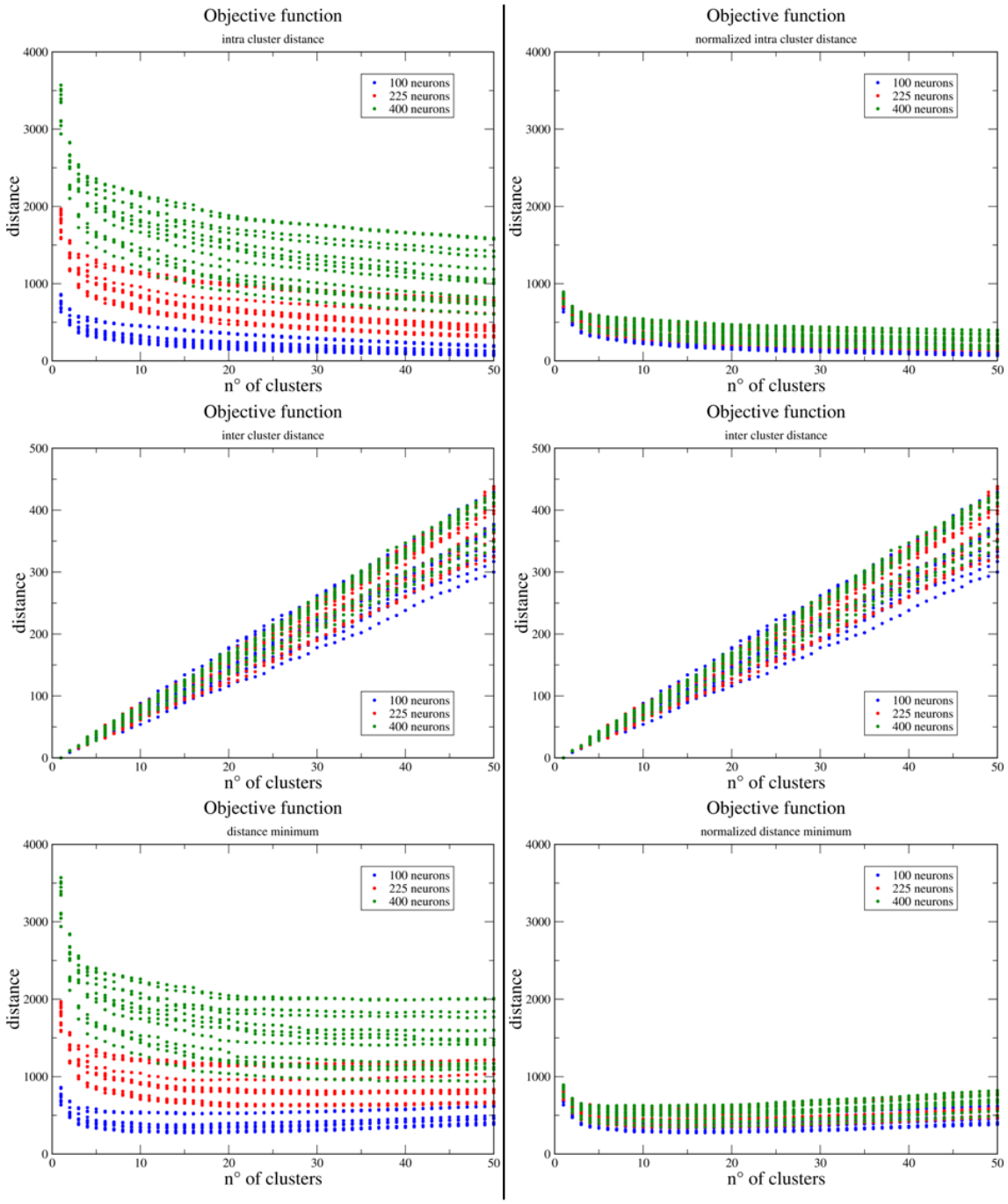



*Figure 2.8 – Objective function: on the left the graphs without the normalization factor, on the right the graphs with the normalization factor. The data are divided with respect to the number of neurons of the experiment (blue = 100, red = 225, green = 400). Each member of the objective function E32 is represented: at top the decreasing intra cluster distance, in the middle the increasing inter cluster distance, at bottom the sum of the terms.*

the experiment. On the contrary, the intra cluster distance, evaluated by the first term of (E32) (at the top of the Figure), is influenced.

The motivation for the choice of the objective function (E32) is twofold; first it has a unique minimum and second it has been empirically shown that it is capable to deal with the subjective assessment of clusters (64).

### *2.3.4 The SOM parameter optimization.*

The numerical experiments are used to fit a linear regression model of the performance measure (E32) as function of the SOM's most relevant parameters. The linear regression model allows to select the optimal values of the SOM's design parameters. The linear regression analysis is performed in two stages: a stepwise regression (65) to select the relevant parameters, and a linear regression model fitting to model the unknown mapping between such parameters and the performance measure (E32).

The Stepwise regression parameters were set to *Prob to Enter* = 0.05, *Prob to Leave* = 0.05 and *Direction* = Mixed and *Rules* = No Rules.

In particular, *Prob to Enter* is the significance probability that must be attributed to a regressor term for it to be considered as a forward step and entered into the model; *Prob to Leave* is the significance probability that must be attributed to a regressor term in order for it to be considered as a backward step and removed from the model; *Direction* lets you choose how you want variables to enter the regression equation and *Rules* can change the rules that are applied when there is a hierarchy of terms in the model. The choice *Direction* = Mixed, alternates the forward and backward steps. It includes the most significant term that satisfies *Prob to Enter* and removes the least significant term satisfying *Prob to Leave*. It continues removing terms until the remaining terms are significant and then it changes to the forward direction. *Rules* = No, gives the selection routine complete freedom to choose terms.

JMP software was used for data analysis and linear regression (66).

### *2.3.5 Optimal sampling rate of MD trajectories.*

When dealing with a nanoseconds MD trajectory it is desirable to train the SOM with a minimum number of selected structures while maintaining a reliable picture of the protein dynamics. To this extent, the effect of the sampling rate on the SOM learning process was assessed and the "minimum" number of frames to extract from a trajectory was identified.

To describe the sampling rate analysis the following notation is required. Let $Q_i^{(n)}$ be the $i^{th}$ random sample for a trajectory consisting of $n$ points (conformations) and $SOM_i^{(n)}$ be the SOM trained on the dataset $Q_i^{(n)}$ with the optimal parameters as determined through the SOM optimization procedure. Furthermore, let $\underline{h_i^{(n)}}$ and $H_i^{(n)}$ be respectively the hits vector of the $SOM_i^{(n)}$ when the dataset $Q_i^{(n)}$ and the dataset $Q$ are submitted. Finally, let $\chi_i^{(n)}$ be the Chi-square statistic computed on non null cells of hits vector $H_i^{(n)}$ and cells of hits vector $\underline{h_i^{(n)}}$. The single time-step analysis procedure, for a given number of points $n$ and for a given number of samples $k$, is depicted in Figure 2.9 and summarized as follows:

1) extract $k$ random samples from the trajectory dataset $Q$;
2) train a SOM for each sample $Q_i^{(n)}$, $i=1, \ldots, k$, to obtain $SOM_i^{(n)}$;
3) assign points in $Q_i^{(n)}$ to the neurons of $SOM_i^{(n)}$ to obtain the hits vector $\underline{h_i^{(n)}}$;
4) assign points in $Q$ to the neurons of $SOM_i^{(n)}$ to obtain the hits vector $H_i^{(n)}$;
5) compute the Chi-squared statistic $\chi_i^{(n)}$, $i=1, \ldots, k$;
6) perform $k$ Cressie-Read Chi-square tests (67) ;
7) if no test is rejected, then accept the hypothesis that a number of sampling points equals to $n$ does not produce a hits vector significantly different from the one obtained when using the full set of data points, i.e. the dataset $Q$.

The single sampling rate analysis procedure is applied to different number of points $n$ to find the minimum value $n^*$ such that the null hypothesis is not rejected.

This value ensures that the clusters extracted from SOM learning are not influenced by the sampling rate applied to the dataset *Q*.

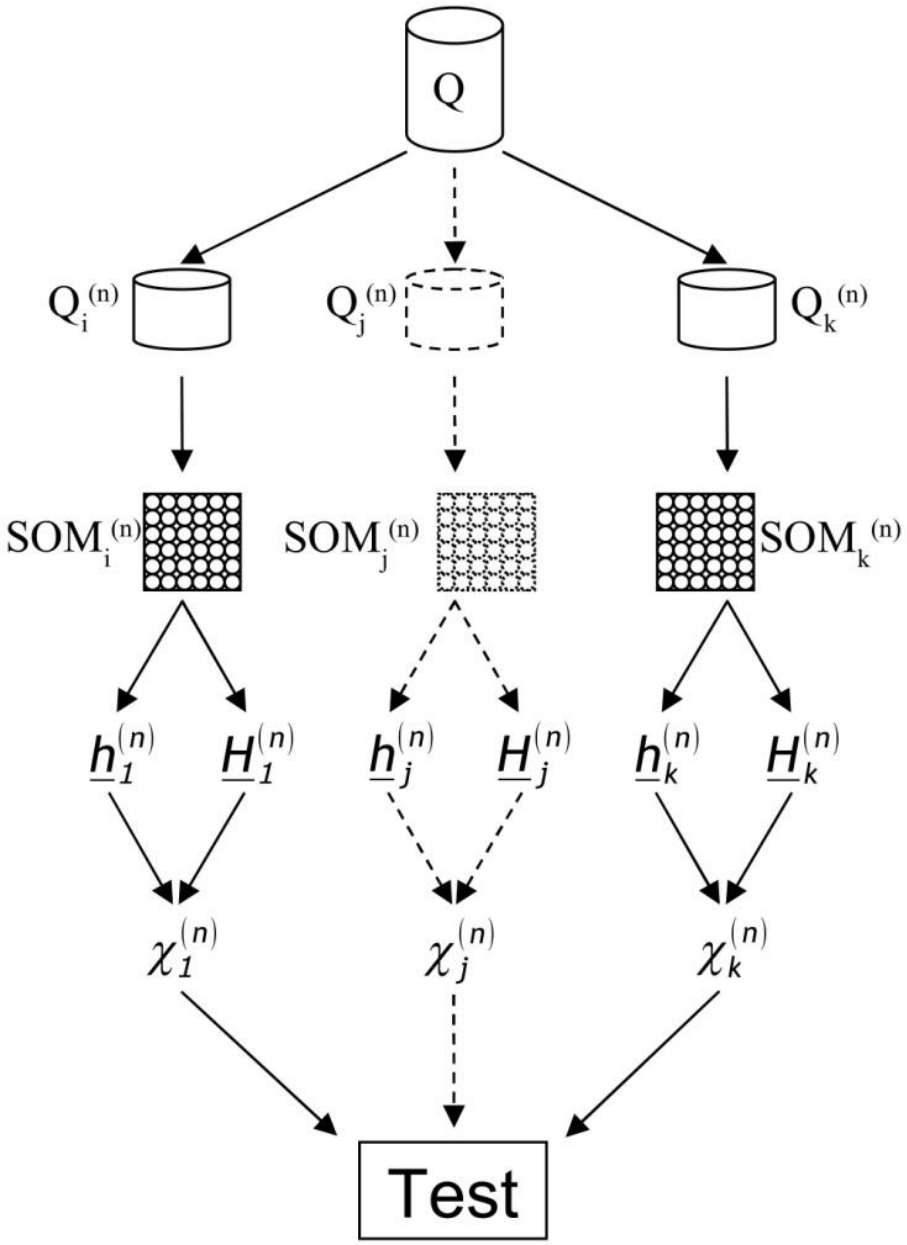



*Figure 2.9 – Optimal sampling rate of MD trajectories: diagram of the procedure to test the reliability of SOM clustering at different sampling rates*

### *2.3.6 SOM clustering for structural ensembles.*

This task is devoted to give an interpretation of the learning process applied to the SOM. The SOM algorithm, when presented with an ensemble of structures, learns the features of the protein dynamics and consequently adapts the weights of its neurons. Therefore the neurons' weights $(\underline{w_i})$ become the coordinates of representative geometries and the problem of performing cluster analysis on the original ensemble is reduced to clustering only the representatives. To this aim we used an algorithm which combines hierarchical complete linkage clustering (68) with the following instance of the Mojena's rule (E33) (69).

$$d^* \geq \overline{d} + 2.75\,\sigma_d \qquad \text{(E33)}$$

Where $\overline{d}$ is the mean of the sequence of distances of the dendrogram $(d_{1,\dots,}d_{100})$, $\sigma_d$ is the standard deviation of $d$, 2.75 is the coefficient applied for the complete linkage and $d^*$ is the value of $d$ that define the optimal number of cluster. The value of $d^*$ is the first value of $(d_{1,\dots,}d_{100})$ which satisfies the inequality in (E33) (see Fig. 2.10).

The procedure can be summarized as follows, taking as reference a SOM of 100 neurons:

1. Run the hierarchical complete linkage clustering on the set $W = \{w_1, \dots, w_{100}\}$ to obtain the dendrogram $D = \{T, (d_{1,\dots,}d_{100})\}$, where $T$ is the clustering tree while $(d_{1,\dots,}d_{100})$ is the corresponding sequence of distance values. By definition $(d_1 < d_2 < \cdots < d_{100})$ where $d_{100}$ represents the value of the maximum distance between weight vectors belonging to the cluster, obtained as the results of the *100th* joining operation. The hierarchical clustering has been performed by using the MATLAB Statistics Toolbox (70);
2. Use the Mojena's rule to find the optimal number of clusters $q^*$, i.e. the optimal number of clusters while $C_i$, *i=1, …,*$q^*$ consists of the set of indices

for the neurons belonging to the $i^{th}$ cluster. Use the tree $T$ together with $q^*$ to compute the optimal segmentation $C^* = \{C_1, \dots, C_{q^*}\}$. The Mojena's rule is applied by using the MATLAB code made available by the courtesy of Prof. Josep Antoni Martín Fernández (http://ima.udg.edu/~jamf/).

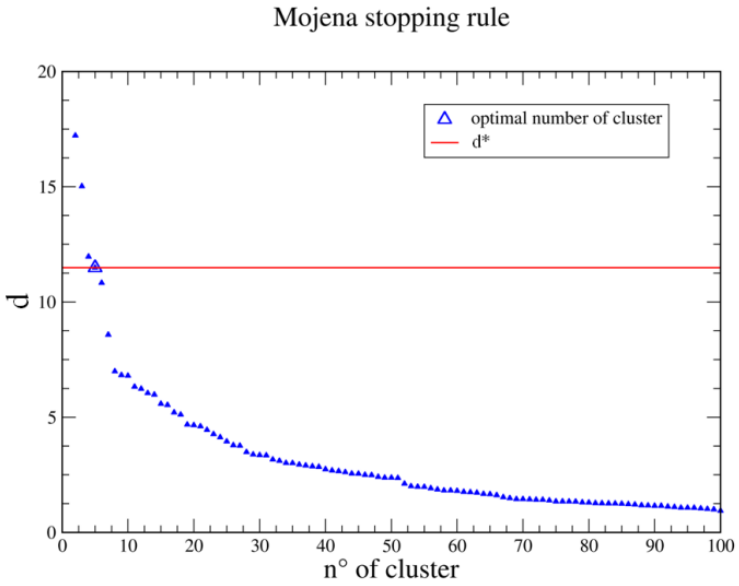



*Figure 2.10 – Mojena stopping rule. In the graph each blue triangle represent the distance for the aggregation of increasing number of clusters and the bigger triangle corresponds to the optimal number of clusters. In red the line corresponding to the cutoff value as defined in E33.*

Once defined the optimal number of clusters it is possible to extract the conformations that belong to each cluster.

Moreover it is possible to extract the centroid of each j$^{th}$ cluster, $\underline{\mu}_j$. The centroid $\underline{\mu}_j$ and the weight vector of each neuron "i", $\underline{w}_{ij}$, are vectors with the same dimension of the input data $(\underline{x}_{t=i})$. As described in Par 2.2.2, the update process of the weight vector $\underline{w}_{ij}$ produces vector that are made similar to the data "won" during this phase. This means that at the end of the learning process these vectors are very similar to the original data but in principle the values that compose this vector can be different from any original data. In our case this means that, often they do not represent real conformations. We can extract the real conformation that best describes each neuron searching for the input data vector closest to $\underline{w}_{ij}$. Similarly

we can define the real conformation closest to the centroid $\underline{\mu}_j$, and we can use this conformation to provide the maximum summarization of the cluster. This means that it is possible to extract information at three different levels:

a) cluster level; i.e. to summarize each cluster with the conformations "won" by its neurons;
b) neuron level; i.e. to summarize each cluster by means of its most representative neuron;
c) centroid level; i.e. to summarize each cluster by using only the conformation that best describes its centroid.

### 2.3.7 *Optimization of the protocol for the study case of the SH3 protein domains*

The SOM protocol was optimized by using the MD trajectories of the Spc-SH3 domain and its six mutants as a study case (see Cap. 3 for the description of the system).

As presented in Par 2.3.1, each sampled conformation was described by the Cartesian coordinates of the Cα atoms. In these study cases, the number of Cα in structurally equivalent positions in all the domains is 55. Therefore, the input data vectors $(\underline{x}_{t=i})$ presented as inputs to the SOM were vectors consisting of 165 elements (55 Cα x 3 Cartesian coordinates).

The experimental plan, defined in Par 2.3.2, consisting of 36 runs with three replicas for each run, was used for the following four datasets; the wild-type SH3 (WT) dynamic, the R21G mutant dynamic, the combined trajectories of WT and R21G (called WT+R21G case), and the combined trajectories of WT and its six mutants (R21A, R21G, N47A, N47G, A56G, A56S, called ALL case). These datasets have been selected to optimize the SOM both for the analysis of a single trajectory and for multiple comparisons of trajectories. Among the SH3 mutants, the ones that cause a relevant increase of flexibility with respect to the wild type have been selected (see Cap. 3 for more details).

Therefore, a total of 432 runs have been performed.

$$36\ \frac{experiments}{dataset} \times 3\ \frac{runs}{experiments} \times 4\ dataset = 432\ runs$$

The dimensions of the matrices given in input to the SOM ($\boldsymbol{X}$) are different for the different datasets. In the analyses of a single trajectory the dimension (row x column) of the matrices is 400 x 165, in the analyses of pair of trajectories is 800 x 165 and in the analyses of the seven trajectories is 2800 x 165.

The response variable is the minimum normalized distance as defined in equation (E32).

The summary of the linear regression model fitting are reported in Table 2.3 and Table 2.4, and the regression curve is reported in Fig. 2.11. The regression model was satisfactory, with similar values of $R^2$ and $R^2_{adj}$ ($R^2$ = 0.937 and $R^2_{adj}$ = 0.936).

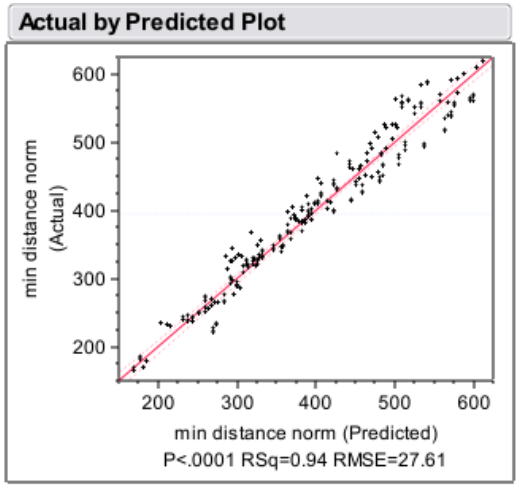



*Figure 2.11 – Regression curve of the experiments: the figure illustrates the actual versus predicted plot of the linear regression model which links the following SOM design parameters: Map size, Radius, Training length and Neighbour function, to the response variable, i.e. the minimum distance normalized (E32).*

*Table 2.3: Linear regression, summary of fit. $R^2$ measures the proportion of the variation around the mean explained by the linear or polynomial model; $R^2_{Adj}$ adjusts the $R^2$ value to make it more comparable over models with different numbers of parameters by using the degrees of freedom in its computation. RMSE (Root Mean Square Error) estimates the standard deviation of the random error; Mean of Response is the sample mean (arithmetic average) of the response variable; Observations is the number of observations used to estimate the fit*

| Parameter | Value |
|---|---|
| $R^2$ | 0.937 |
| $R^2_{adj}$ | 0.936 |
| RMSE | 27.61 |
| Mean of Response | 395.01 |
| Observations | 432 |

*Table 2.4: Linear regression, effect tests. Source, lists the names of the effects in the model; Nparm is the number of parameters associated with the effect; DF is the degrees of freedom for the effect test; Sum of Squares is the sum of squares for the hypothesis that the listed effect is zero; F Ratio is the F-statistic for testing that the effect is zero; Prob>F is the significance probability for the F-ratio.*

| Source | Nparm | DF | Sum of Squares | F Ratio | Prob > F |
|---|---|---|---|---|---|
| Map size | 1 | 1 | 1,022,810.9 | 1,341.853 | <0.0001 |
| Radius | 1 | 1 | 39,779.1 | 52.187 | <0.0001 |
| Training length | 1 | 1 | 17,056.8 | 22.377 | <0.0001 |
| Neighbor function | 2 | 2 | 2,935,552.0 | 1,925.615 | <0.0001 |
| TYPEMOL | 3 | 3 | 805,604.4 | 352.298 | <0.0001 |

The following design parameters have been judged to have statistical significance at the p-value cutoff of 0.01: Map size, Radius, Training length, Neighbour function and TYPEMOL (see Table 2.4), the latter being a categorical variable, which can take the following values: WT, R21G, WT+R21G and ALL.

The optimal settings of the SOM design parameters, for each value of the categorical variable TYPEMOL, were: Map size=100, Radius=3, Training length=5,000 and Neighbour function=gaussian. Table 2.5 reports the results of a validation test on the regression model. The difference between the predicted and the actual optimal values is small, confirming that the regression model can reliably predict the value of the performance measure (E32).

Interestingly the parameters found to be the only relevant for the analysis of this specific kind of data are the same in the four datasets studied and, more

interestingly, in the four cases the optimal value of each parameter is the same. This is not so straightforward by considering the different nature of the four datasets. They are different not only in terms of their dimensions (from 400 to 2,800 vectors) but, more relevant, in terms of amplitude of the conformational changes. Even if it is not possible to state that the value of these parameters will be optimal for the analyses of all the possible datasets composed by data extracted by the MD simulations, we can say that these are the most relevant parameters that have to be optimized for the considered cases.

*Table 2.5: Optimal SOM's design parameter values. For each parameter the value that optimizes the objective function is reported. Using these parameters: Actual is the result of the experiment performed; Predicted is the result obtained using the regression model.*

| Case | Map size | Radius | Training length | Neighbor function | Actual | Predicted |
|---|---|---|---|---|---|---|
| WT | 100 | 3 | 5,000 | gaussian | 158 | 156 |
| R21G | 100 | 3 | 5,000 | gaussian | 265 | 270 |
| WT+R21G | 100 | 3 | 5,000 | gaussian | 259 | 238 |
| ALL | 100 | 3 | 5,000 | gaussian | 254 | 246 |

Each system was simulated for 40 ns (see Chapter 3). At first, a sampling rate equal to 0.01 was set, i.e. out of the available 40,000 trajectory points (one each ps), 400 were randomly selected to describe each trajectory. A different sampling rate could have been resulted in a different trajectory clustering. Therefore, the appropriateness of the sampling rate was checked by the method described in Par. 2.3.5. The analysis was performed with three replicas for each sampling rate: 1/2 ps, 1/4 ps, 1/8 ps, 1/16 ps, 1/25 ps, 1/50 ps, 1/100 ps, 1/500 ps and 1/1,000 ps (corresponding to sampling size from 20,000 to 40 conformations): 20,000, 10,000, 5,000, 2,500, 1,600, 800, 400, 80, and 40 conformations. Three cases were selected to study the different sampling rate: the WT, an example of a trajectory without relevant conformational changes, the R21G mutant, the case with the highest conformational changes during the simulation, the N47A mutant, to describe an intermediate situation.

The results, reported in Table 2.6, show that the null hypothesis is not rejected for samples larger than 400 conformations. Thus, at a rate of 1/100 ps or greater the results of the SOM analysis are not influenced by the sampling rate. It means that using this sampling rate a complete picture of the dynamics, in terms of representation of the conformations sampled, is obtained.

The optimal SOM obtained for each dataset, consisting of *100* neurons, is summarized by the set of neuron's weights $W = \{w_1, \dots, w_{100}\}$, where each $\underline{w_i}$ is a vector with the same dimension of the input pattern $(\underline{x}_{t=i})$.
After application of the cluster analysis algorithm described in Par 2.3.6 to the obtained neurons, information useful for conformational and functional analysis of the SH3 domains were extracted from the maps of both the single and the combined trajectories. Results will be presented and discussed in Chapter 3.

*Table 2.6: Test on the effects of different sampling rates for the WT SH3 and the R21G and N47A mutants: Sampling rate analysis for the WT SH3 and the R21G and N47A mutants. Statistic, p-value and the output of the hypothesis test are reported; (-) means that the null hypothesis cannot be rejected while (*) means that the null hypothesis is rejected.*

| | | **statistic** | | | **p-value** | | | **result** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **frequency** | **sample** | **WT** | **R21G** | **N47A** | **WT** | **R21G** | **N47A** | **WT** | **R21G** | **N47A** |
| 20,000 | 1 | 25 | 27 | 16 | 1.00 | 1.00 | 1.00 | - | - | - |
| | 2 | 17 | 28 | 17 | 1.00 | 1.00 | 1.00 | - | - | - |
| | 3 | 19 | 29 | 22 | 1.00 | 1.00 | 1.00 | - | - | - |
| 10,000 | 1 | 29 | 35 | 34 | 1.00 | 1.00 | 0.98 | - | - | - |
| | 2 | 28 | 37 | 36 | 1.00 | 1.00 | 0.98 | - | - | - |
| | 3 | 33 | 36 | 21 | 0.99 | 1.00 | 1.00 | - | - | - |
| 5,000 | 1 | 30 | 33 | 30 | 1.00 | 1.00 | 0.98 | - | - | - |
| | 2 | 49 | 49 | 28 | 0.91 | 0.98 | 0.98 | - | - | - |
| | 3 | 34 | 37 | 30 | 0.99 | 1.00 | 0.99 | - | - | - |
| 2,500 | 1 | 63 | 53 | 23 | 0.26 | 0.92 | 0.99 | - | - | - |
| | 2 | 51 | 48 | 56 | 0.76 | 0.99 | 0.22 | - | - | - |
| | 3 | 39 | 44 | 34 | 0.92 | 0.99 | 0.98 | - | - | - |
| 1,600 | 1 | 43 | 49 | 31 | 0.82 | 0.93 | 0.96 | - | - | - |
| | 2 | 42 | 63 | 28 | 0.86 | 0.74 | 0.95 | - | - | - |
| | 3 | 45 | 48 | 41 | 0.90 | 0.97 | 0.77 | - | - | - |
| 800 | 1 | 40 | 57 | 73 | 0.92 | 0.89 | 0.06 | - | - | - |
| | 2 | 57 | 43 | 49 | 0.52 | 0.99 | 0.47 | - | - | - |
| | 3 | 54 | 44 | 36 | 0.48 | 0.98 | 0.80 | - | - | - |
| 400 | 1 | 48 | 73 | 39 | 0.63 | 0.11 | 0.49 | - | - | - |
| | 2 | 50 | 49 | 55 | 0.55 | 0.93 | 0.16 | - | - | - |
| | 3 | 34 | 39 | 46 | 0.96 | 0.99 | 0.47 | - | - | - |
| 80 | 1 | 40 | 83 | 54 | 0.28 | 0.00 | 0.02 | - | * | * |
| | 2 | 63 | 64 | 38 | 0.00 | 0.00 | 0.08 | * | * | - |
| | 3 | 81 | 46 | 50 | 0.00 | 0.43 | 0.04 | * | - | * |
| 40 | 1 | 167 | 84 | 67 | 0.00 | 0.00 | 0.00 | * | * | * |
| | 2 | 105 | 62 | 95 | 0.00 | 0.00 | 0.00 | * | * | * |
| | 3 | 101 | 60 | 67 | 0.00 | 0.00 | 0.00 | * | * | * |

# Chapter 3

## MODULATION OF FLEXIBILITY BY MUTAGENESIS: SPC-SH3 DOMAIN MUTANTS

### 3.1 Selection of the study case

As anticipated in Chapter 1, we were interested in studying the modulation of flexibility given by mutations, and its effects on ligand binding.
The study case selected for this analysis is composed by the α-spectrin SH3 (Spc-SH3) domain and a group of its single-site mutants. The MD trajectories of this group of mutants were also used for the optimization of the SOM protocol for the analysis of structural ensembles (59), reported in Chapter 2.

As shown in Figure 3.1a, the crystal structure of the wild-type Spc-SH3 domain (62 residues, PDB code: 1SHG) is characterized by five antiparallel β-strands that form two orthogonal β-sheets. A long 19-residue loop that includes three isolated β-bridges (RT loop), connects the first two strands, while two loops (commonly termed n-src and distal loop) connect the β2 - β3 and the β3 - β4 strands, respectively, and a short $3_{10}$ helix joins the β4 and β5 strands (71).

The Spc-SH3 domain binds the decapeptide APSYSPPPPP (p41), although with moderate affinity ($K_d$ = 83 ± 7 μM) (38). Proline-rich polypeptides usually bind SH3 domains in a polyproline II (PPII) helical conformation, and the typical SH3 binding surface comprises two hydrophobic grooves lined mainly by aromatic residues, and a specificity pocket flanked by the RT and n-src loops (72; 73). Several studies have demonstrated that the conformational dynamics of these loops plays an important role in determining the binding specificity (74). The structure of the R21A Spc-SH3:p41 complex (PDB code: 2JMA, (75)) was obtained by solution NMR and HADDOCK simulations (76). As the R21A mutant is

structurally very similar to the wild-type Spc-SH3 (Figure 3.1b), the structure of this complex confirms that the binding mode of p41 reproduces the general features of the SH3 domains' binding. The group of residues that generate hydrophobic contacts and hydrogen bonds with p41 (75) are highlighted in the figure. The core SH3 interaction surface is formed by aromatic residues that interact with proline and hydrophobic residues of the ligand, while the specificity pocket is constituted by residues in the RT and n-src loops and in the β4 strand.

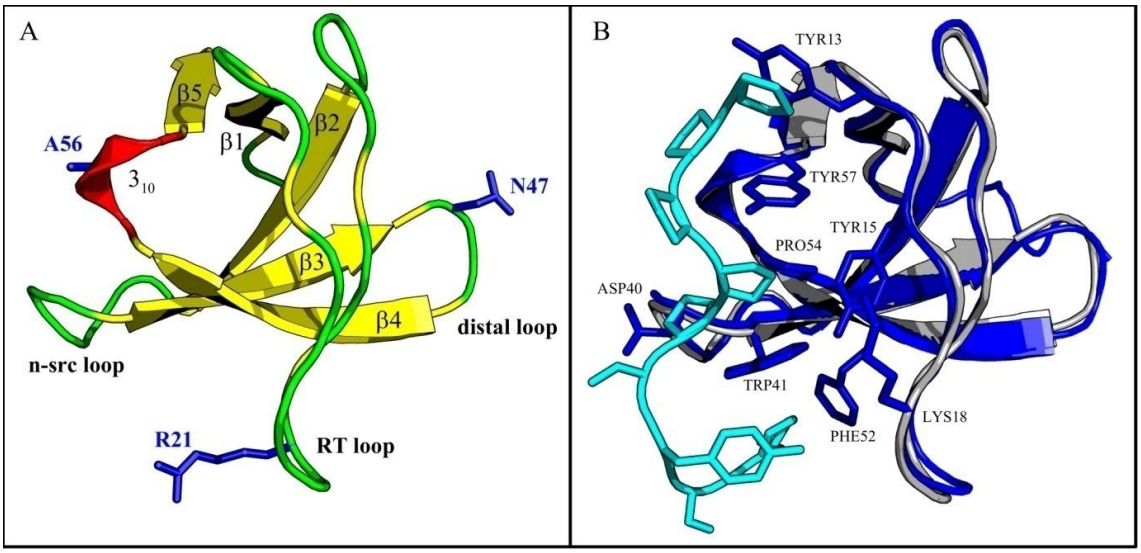



*Figure 3.1: Cartoon representations of the Spc-SH3 domain. A) X-ray structure of WT Spc-SH3 (PDB code: 1SHG). Secondary structures are attributed according to the DSSP program (β-strands: yellow; $3_{10}$ helix: red; loops: green) and labeled according to the generally adopted nomenclature for SH3 domains. The three mutated residues are shown as blue sticks. B) Structure of the R21A Spc-SH3:p41 complex (PDB code: 2JMA), in blue, superimposed to the WT Spc-SH3 structure in grey. Residues that interact with the ligand are shown as sticks. The structure of p41 is shown in cyan*

The six mutants of the Spc-SH3 domain here analysed (and their PDB codes) are: R21A, R21G, A56G and A56S (2F2W, 2F2X, 2F2V and 2CDT, (38)), N47A and N47G (1QKX, 1QKW, (77)). These mutants had been designed to explore how the local perturbations produced by single-point mutations affect both the stability and the p41 affinity of this domain (38).

In particular, it was observed that both mutations of R21, at the tip of the RT loop that flanks the binding specificity pocket (Figure 3.1a), had favourable effects on the p41 affinity, due to the replacement of the bulky arginine side chain by a small side chain. However, the change in affinity was smaller for the R21G mutant, and it was hypothesized that this may be caused by the increased conformational freedom of the RT loop. Both mutations of N47, in the distal loop (about 10 Å from the binding site, Figure 3.1a), produced significant changes in the stability of the domain and a reduction of p41 affinity, which suggested the idea of a cooperative pathway between the distal loop and the binding site (78). This was supported by the identification of a stabilizing salt-bridge between the distal and the RT loops (R49 and E17) in the wild-type X-ray structure, that was not evidenced in both the mutants in the N47 position (78). Finally, while mutation of A56, in the $3_{10}$ helix belonging to one of the hydrophobic binding grooves (Figure 3.1a), to Ser did not alter significantly the p41 affinity, the mutation to Gly produced a great reduction in affinity, that was attributed to the increase of conformational freedom induced by this mutation.

## *3.2 MD simulations of the SH3 mutants*

### *3.2.1 Simulation protocol*

The atomistic dynamics of the set of proteins under study was simulated using the GROMACS package (version 3.3.3) (39; 42; 43) with the GROMOS96 43a2 force field, by following the general protocol described in Par. 2.1.2 and 2.1.3. The specific details set for this group of domains is outlined in the following.

All structures were inserted in an octahedral box with explicit solvent and simulated with periodic boundary conditions. Water molecules were described by a simple point charge (SPC) model (49) and the box size was set to ensure a distance of at least 1.2 nm between the protein and the box boundaries. The solvent was relaxed with a 5 ps MD simulation, then the systems were neutralized by

insertion of counter ions, and a short minimization with steepest descent was performed up to convergence on maximum force lower than 1000 kJ/mol nm. The resulting systems were simulated for 40 ns in the NPT ensemble. Long-range electrostatic interactions were calculated with the particle mesh Ewald (PME) summation method (40). A thermal bath was independently coupled with protein and solvent using a Berendsen thermostat at 300 K with coupling period of 0.1 ps. The internal degrees of freedom of water were constrained by the Settle algorithm (50), while all bond distances in the protein were constrained by the LINCS algorithm (52). The integration step was set to 2 fs.

The conformational dynamics of the X-ray structures of the wild type (WT) and the six mutated Spc-SH3 domains were analysed by 40 ns MD simulations.

### *3.2.2 Analysis of the trajectories*

The analysis of the root mean square deviation (RMSD) to the starting structure confirmed a general stability of all trajectories, with an equilibration time around 1 ns and a temperature around 300K during the simulation (see Fig. 3.2). As shown in Figure 3.2 different behaviours are observed during the simulations. The dynamics of the WT SH3 (Fig 3.2a) and of the R21A mutant (Fig 3.2b) show low flexibility of the domains, with an average RMSD of 0.10 nm, and are characterized by the absence of relevant conformational transitions. All the other simulations have higher average values of RMSD and show significant transitions. The R21G mutant (Fig. 3.2b) shows a significant conformational change between 15 and 30 ns, with average RMSD of 0.25 nm. In the other mutants the transitions are shorter and the RMSD has average values of about 0.20 nm.

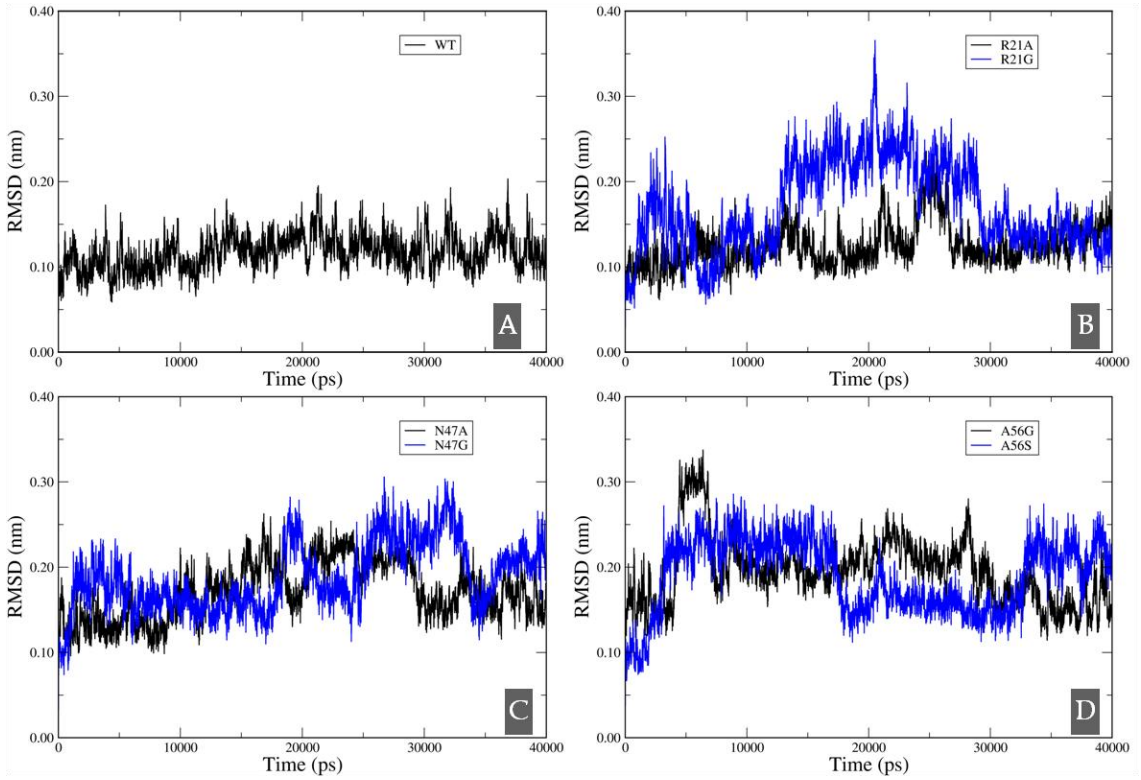



*Figure 3.2: RMSD computed on Ca for the wild-type SH3 and the six mutants.*

A test on the completeness of conformational sampling was performed by using the overlap between the conformational space spanned in each half of the simulation and that of the overall simulation as a convergence index, as described in Par. 2.1.4. The results, reported in Table 3.1, clearly indicate an overlap of more than 70% with the whole trajectory in both the halves of the simulation for all the systems, suggesting a good convergence of sampling.

The most informative directions of motion were extracted by Essential Dynamics analysis (see Par. 2.1.5) and the fraction of total motion described by different subspaces was evaluated to identify the extent of the essential space for each domain. Results reported in Table 3.2 indicate that in all cases more than 79% of the motion is described by the first 20 eigendirections (corresponding to 12% of the total space), while the first 30 directions (18% of total space) can explain more than 85% of the conformational flexibility of all domains. Therefore a 30 dimensional essential space was selected for the following analyses.

*Table 3.1: Overlap of sampling in the MD simulations of the SH3 domains. Values represents the overlap between the conformational space spanned by each half of the simulation and that of the overall trajectory.*

| | 1 - 20 ns | 20 - 40 ns |
|---|---|---|
| WT | 0.87 | 0.86 |
| R21A | 0.76 | 0.77 |
| R21G | 0.82 | 0.78 |
| N47A | 0.70 | 0.75 |
| N47G | 0.72 | 0.78 |
| A56G | 0.71 | 0.71 |
| A56S | 0.83 | 0.80 |

*Table 3.2: Distribution of motion in different subspaces for each MD simulation. Values refer to the percentage of total space described by the eigenvectors.*

| Eigenvectors | WT | R21A | R21G | N47A | N47G | A56G | A56S |
|---|---|---|---|---|---|---|---|
| 1 - 10 | 65.2 | 69.9 | 81.5 | 73.5 | 77.1 | 80.0 | 80.3 |
| 1 - 20 | 79.3 | 82.1 | 89.4 | 85.2 | 87.0 | 88.8 | 88.6 |
| 1 - 30 | 86.3 | 88.0 | 92.8 | 90.4 | 91.4 | 92.5 | 92.3 |
| 1 - 60 | 95.0 | 95.5 | 97.3 | 96.5 | 96.7 | 97.1 | 97.1 |
| 1 - 165 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

The comparison of the local flexibility on a residue base was performed by using the RMSF, a traditional index for MD simulation analysis (Par. 2.1.6). A set of comparative plots of the RMSF on the positions of the Cα atoms in the essential space is shown in Figure 3.3, where only equivalent residues in the preliminary structure-based alignment are included and the secondary structures are reported in the bottom part of each graph for reference.

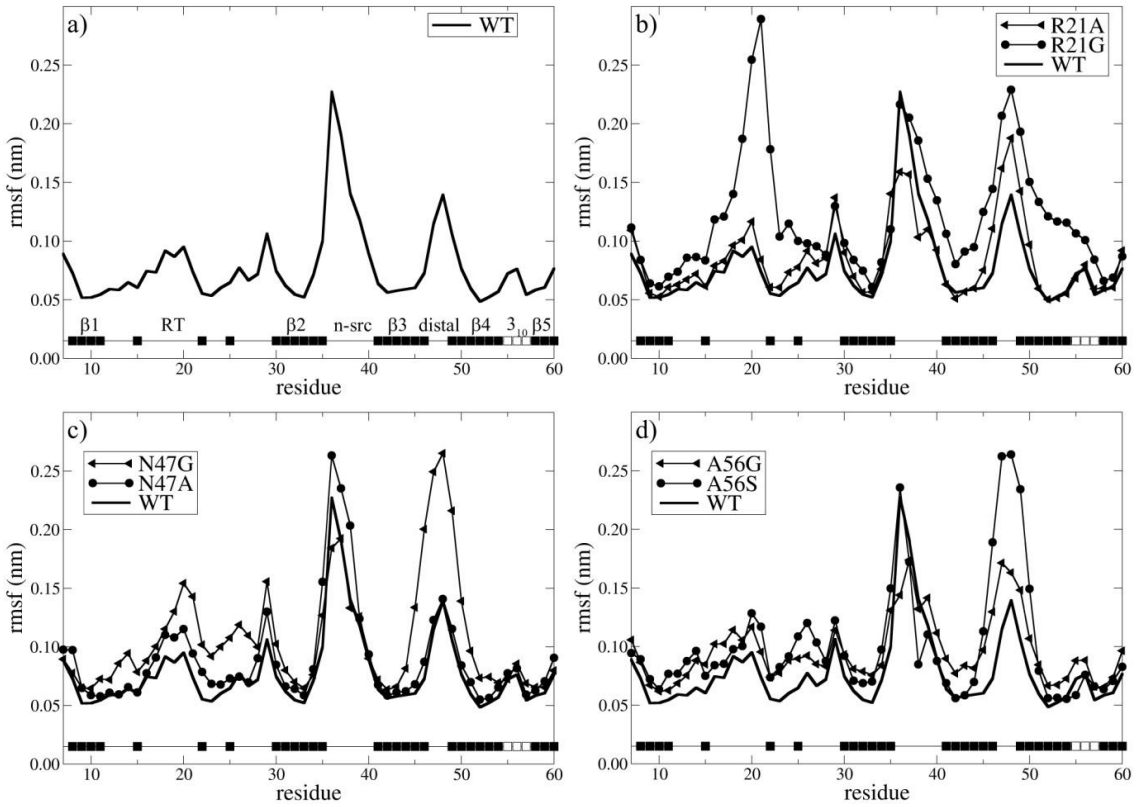



*Figure 3.3 - Plot of RMSF versus residue position in the essential space. MD simulations of: a) WT SH3; b) R21A and R21G mutants, compared to the WT; c) N47G and N47A mutants, compared to the WT; d) A56G and A56S mutants, compared to the WT. Residue numbers are modified according to the structure-based alignment. Secondary structures are attributed according to the DSSP program (β-strands: black squares; $3_{10}$ helix: white square) and labeled according to the nomenclature generally adopted for SH3 domains*

The WT SH3 structure is rather constrained (fluctuations around 0.1 nm), with only a region of flexibility in the n-src and the distal loops (Fig. 3.3a), while the effects of single-point mutations in each of the three positions, R21, N47 and A56 (see Fig. 3.1a), can be clearly observed in Figures 3.3b, c and d, where the RMSF of each group of mutants is superimposed on the WT plot.

The R21A mutation does not affect significantly the overall domain flexibility, with only a slight increase in the distal loop peak. On the contrary, the R21G mutation causes a relevant increase of flexibility, mainly in the long RT loop (with the highest fluctuation around 0.3 nm at tip of the loop, near the mutated residue) and

in the region including β3, the distal loop, β4 and the helix $3_{10}$ (Figure 3.3b). This result confirms the hypothesis on the role of conformational flexibility in reducing the binding affinity for this mutation (38).

The mutation to Ala at position 47 has little effect on flexibility, while a change to Gly in the same position causes an increased flexibility of all the region around this position (distal loop and nearby strands) as well as in the whole RT loop (Figure 3.3c). These last findings confirm the hypothesis that the N47G mutation causes a reduction in the p41 affinity through a long-range propagation of the local perturbation to the binding site (78).

Differently from the previous cases, a mutation to Gly in position 56, in the $3_{10}$ helix, does not alter significantly either the local or the global flexibility of the domain; an increase of flexibility in the distal loop region is observed for the A56S mutant (Figure 3.3d). From these results, the conformational freedom of these two mutants does not seem to be related to the p41 binding affinity (38).

### *3.3 Conformational and functional analysis by SOM clustering*

In this paragraph the role of flexibility of the SH3 domain in binding the p41 decapeptide is analyzed.

The SOM approach developed in this thesis (see Par. 2.3) was applied for clustering both the single trajectories of the wild type and the SH3 mutants and multiple trajectories of different mutants. The WT and R21G trajectories are presented as examples of single trajectory analysis. For describing the analysis of multiple trajectories, the map obtained by combining the trajectories of WT and R21G (WT+R21G) as well as that obtained by the trajectories of the seven mutants (ALL) are presented.

From a methodological point of view, the SOM analysis of these trajectories allowed us to answer three questions about the use of the SOM protocol described

in Par. 2.3: 1) are the SOMs obtained using the optimized parameters able to cluster both trajectories with high conformational fluctuations and trajectories with slight fluctuations, and is it possible to cluster data obtained from different trajectories, characterized by different fluctuations, using one map? 2) are the Cα coordinates a good descriptor in both the above cases? 3) once defined the ability of the map to cluster these data, are the obtained clusters meaningful from a functional point of view?

### *3.3.1 SOMs of single trajectories: R21G, WT*

The map of the R21G conformational ensemble is shown in Figure 3.4 as an example of analysis of a single trajectory where the domain shows the largest conformational flexibility (see Figure 3.3b). Each hexagon of the map represents a neuron and the black area is proportional to the number of hits (classified conformations). Four clusters (Figure 3.4A) were extracted applying the Mojena's rule (69) after hierarchical clustering. The ensembles of conformations in each cluster are shown as ribbon in Figure 3.4B, superimposed on the X-ray structure of the WT SH3 (in black). Cluster 1 (green) contains a large group of representative conformations with limited fluctuations mostly localised in the n-src loop. Clusters 2 (blue) describes a small displacement of the distal loop towards the RT loop, while cluster 3 (violet) comprises conformations with a more extended motion of the same loop, a consequent perturbation of the n-src loop, and a motion of the RT loop tip back towards the distal loop. Conformations in cluster 4 (red) greatly deviate from the WT structure, with a concerted closure motion of the distal and RT loops.

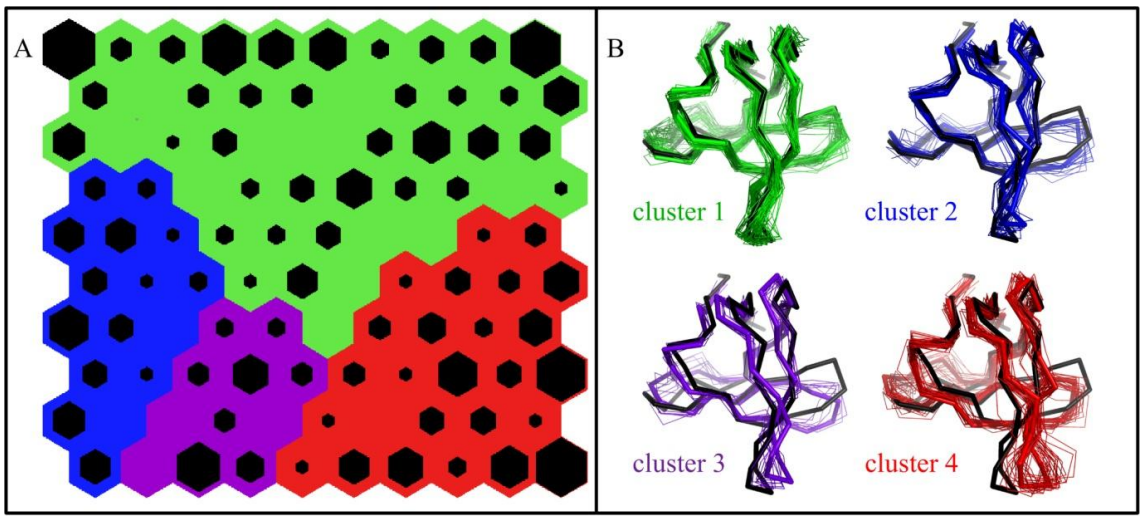



*Figure 3.4 - SOM analysis of the R21G mutant dynamics. A) Self Organizing Map: the number of hits in each neuron are indicated by the black hexagon area; the four clusters obtained by hierarchical clustering of the neurons are indicated by different colors. B) Ribbon representation of the ensembles of hit conformations in each cluster; the representative conformation (see text) is highlighted by a larger ribbon; the X-ray structure of the WT SH3 at t=0 at the trajectory is reported in black for comparison*

As an opposite example, i.e. the analysis of a single trajectory where the domain shows a reduced conformational flexibility, the map of the WT SH3 ensemble is shown in Figure 3.5. Also in this case four clusters (Figure 3.5A) were extracted applying the Mojena's rule (69) after hierarchical clustering. The ensembles of conformations in each cluster are shown as ribbon in Figure 3.5B, superimposed to the X-ray structure (in black). As indicated from the RMSF plot (Fig 3.3a), the most flexible zones are the n-src and the distal loops. The four clusters describe little fluctuations around the equilibrium position, but each cluster slightly differs from the others: cluster 3, in purple in Figure 3.5B, describes conformations close to the starting position of the simulation; cluster 2, in blue, describes little fluctuations of the distal loop; both in cluster 1 (green) and in cluster 4 (red) fluctuations involve the distal and the n-src loops.

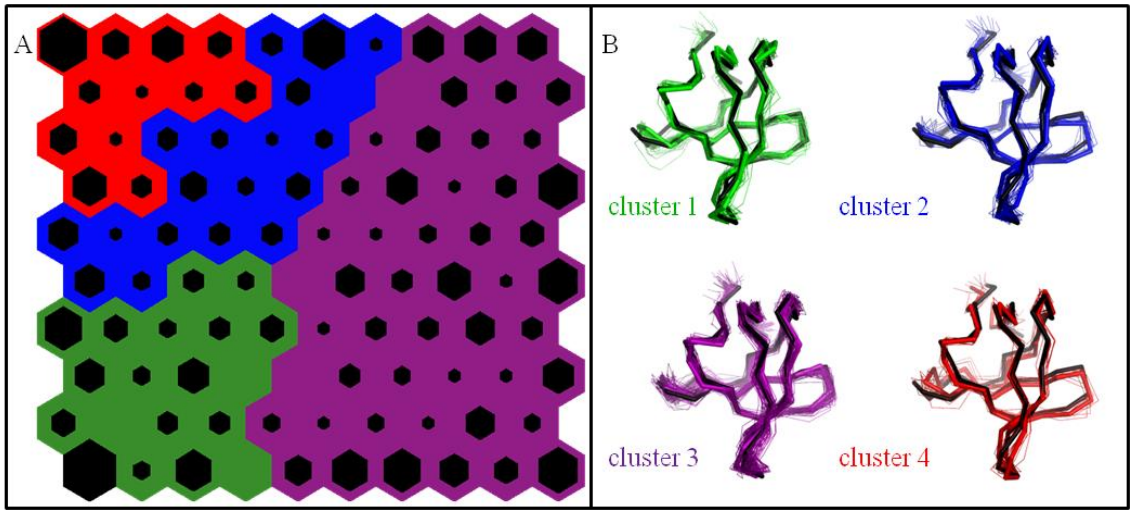


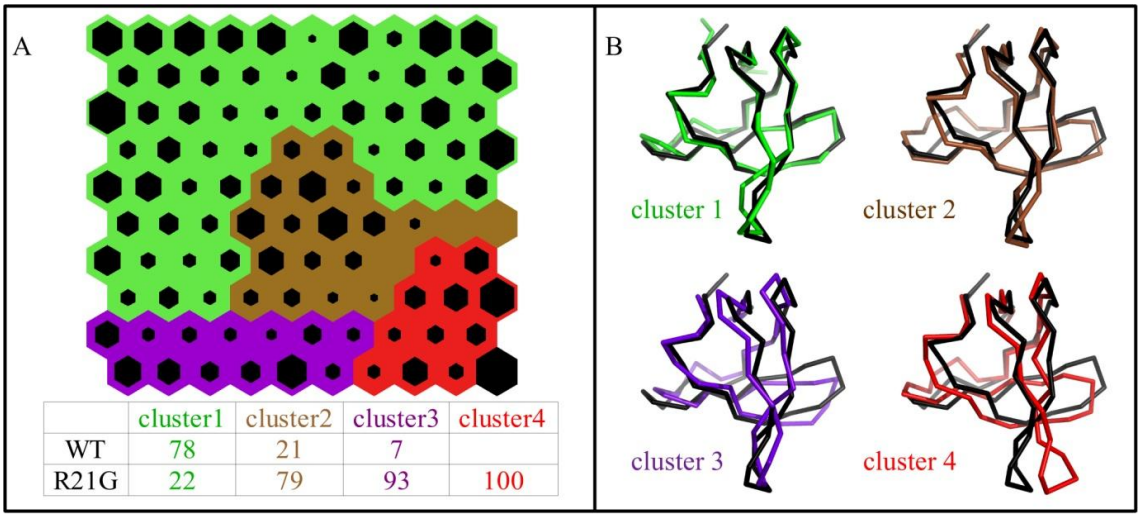

*Figure 3.5 - SOM analysis of the WT SH3 dynamics. A) Self Organizing Map (see Figure 3.4 for the details). B) Ribbon representation of the ensembles of hit conformations in each cluster; the representative conformation (see text) is highlighted by a larger ribbon; the X-ray structure is reported in black for comparison.*

### *3.3.2 SOM of a pair of trajectories: WT and R21G*

Once analysed all the single trajectories of the WT and its mutants, the subsequent step was the analysis of two combined simulations to detect similarities and differences. The results for the WT SH3 and the R21G mutant are discussed. As previously shown, the cluster analysis of the two single trajectories produced clusters with different meaning. In the R21G map the clusters described large conformational changes and fluctuations around the equilibrium position, in that of the WT the cluster described only fluctuations around the equilibrium position.

Four clusters of neurons were extracted from the map trained on the WT and R21G sets of conformations. A table of cluster compositions is reported, along with the map, in Figure 3.6A and the representative conformation of each cluster (i.e. the representative of the neuron that is nearest to the centroid) is shown in Figure 3.6B. The WT ensemble, characterized by low fluctuations, contributes to 78% of cluster 1, that includes conformations similar to its equilibrium structure, and part of cluster 2, whose conformations present small fluctuations in the distal and n-src

loops. On the contrary, cluster 3 and 4 are almost exclusively populated by conformations from the R21G trajectory. Cluster 3 describes more extended fluctuations in the distal and n-src loops and in part of the RT loop, and cluster 4 large concerted motions in the distal and the faced RT loop.



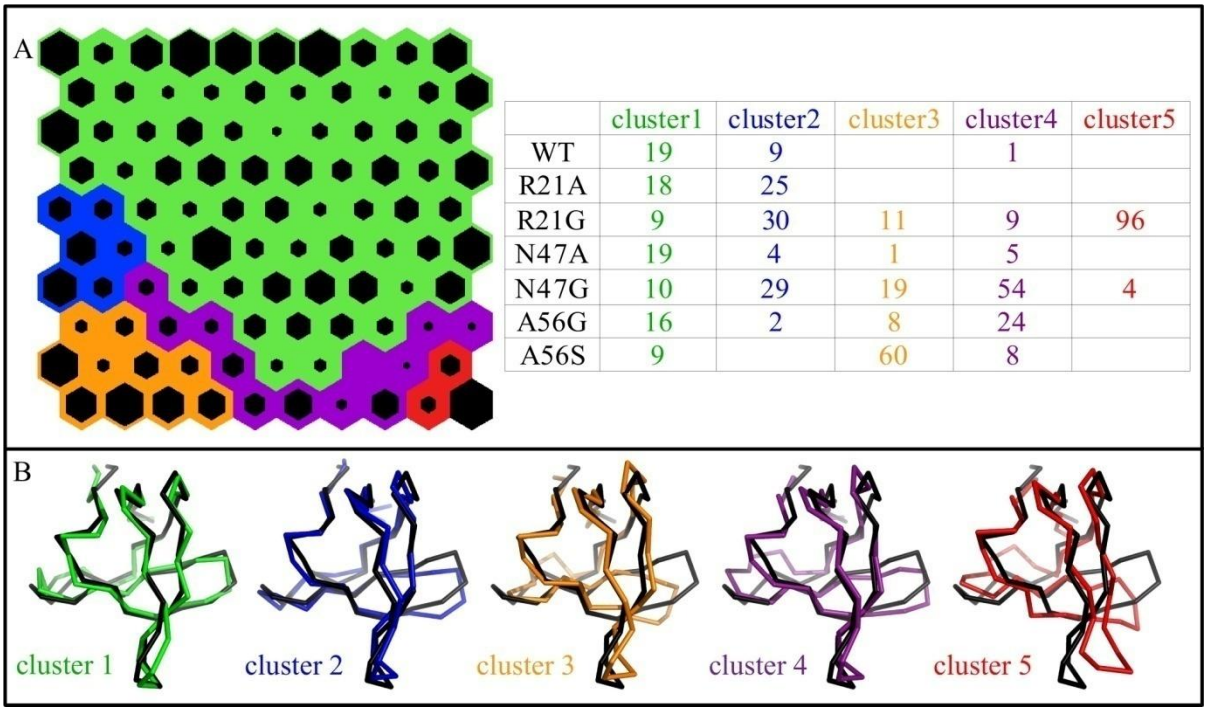

| | cluster1 | cluster2 | cluster3 | cluster4 |
|---|---|---|---|---|
| WT | 78 | 21 | 7 | |
| R21G | 22 | 79 | 93 | 100 |

*Figure 3.6 - SOM analysis for the WT SH3 and the R21G mutant dynamics. A) Self Organizing Map (see Figure 3.4 for the details); the percentage distribution of conformations of each domain in the clusters is reported in the table. B) Ribbon representation of the representative conformation in each cluster, superimposed to the X-ray structure of the WT SH3 (in black).*

### *3.3.3 SOM of WT and the six mutants*

The final step was the analysis of the whole set of trajectories. The clusters obtained with this analysis were studied to describe the functional differences of the mutants correlated with their flexibility.

The SOM of the whole set of ensembles and the resulting five clusters are shown in Figure 3.7. The representative conformations of cluster 1 and 5 (Figure 3.7B) closely resemble those of cluster 1 (low fluctuations) and 4 (large concerted motion of the distal and RT loops) also observed in the SOM of WT and R21G (Figure 3.6B). Clusters 2, 3 and 4 describe intermediate situations with a moderate flexibility of the distal loop (cluster 2) or a large flexibility of the same loop associated with medium to high perturbation of the n-src and the RT loops (clusters 3 and 4).




| | cluster1 | cluster2 | cluster3 | cluster4 | cluster5 |
|---|---|---|---|---|---|
| WT | 19 | 9 | | 1 | |
| R21A | 18 | 25 | | | |
| R21G | 9 | 30 | 11 | 9 | 96 |
| N47A | 19 | 4 | 1 | 5 | |
| N47G | 10 | 29 | 19 | 54 | 4 |
| A56G | 16 | 2 | 8 | 24 | |
| A56S | 9 | | 60 | 8 | |

*Figure 3.7 - SOM analysis of the dynamics of the WT SH3 and the six mutants. A) Self Organizing Map (see Figure 3.4 for the details); the percentage distribution of conformations of each domain in the clusters is reported in the table. B) Ribbon representation of the representative conformation in each cluster, superimposed to the X-ray structure of the WT SH3 (in black).*

To complement and confirm this analysis, the dRMSD (Par 2.1.6) between the average distances of four selected points in the conformational ensemble of each cluster and in the X-ray structure of the WT SH3 was calculated. The selected points (see Figure 3.8) are the Cα atoms of a constrained residue at the N-term of the RT loop (A = L12), and three residues in the most flexible regions of the protein (B = S36 in the n-src loop, C = D48 in the distal loop, D = P20 at the tip of the RT loop).

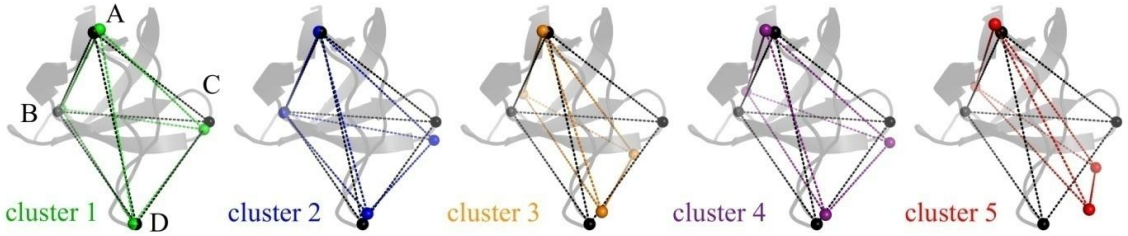



*Figure 3.8 - Distances (dRMSD) among four selected points in each cluster. Dotted lines are colored according to clustering reported in Figure 3.7. The same distances in the WT SH3 structure are reported in black dotted lines. The X-ray structure of the WT SH3, taken as a reference, is represented by grey cartoons and superimposed onto each graph. Points A, B, C, D, are the Cα atom positions in the representative conformation of each cluster of the following residues: A = L12; B = S36 in the n-src loop; C = D48 in the distal loop; D = P20 at the tip of the RT loop.*

*Table 3.3: Distances (dRMSD) among four selected points in each cluster: dRMSD values between points A, B, C, D, in Figure 3.8. The dRMSD are calculated as the average distance of the selected points in the conformational ensemble of each cluster and the same distance in the WT SH3 structure.*

| | dRMSD (Å) | | | |
|---|---|---|---|---|
| | ABD | ACD | ABC | ABCD |
| Cluster 1 | 1.7 | 0.9 | 1.4 | 1.4 |
| Cluster 2 | 1.0 | 1.5 | 1.6 | 1.4 |
| Cluster 3 | 2.2 | 2.3 | 2.7 | 2.4 |
| Cluster 4 | 2.0 | 2.1 | 1.6 | 2.2 |
| Cluster 5 | 3.8 | 2.9 | 2.4 | 3.5 |

Both the visual analysis of the inter-point distances in Figure 3.8 and the ABCD dRMSD values (Table 3.3) indicate that clusters 1 and 2 slightly deviate from the WT structure, where clusters 3 and 4 have more relevant deformations (around 2 Å) and cluster 5 departs more than 3 Å from the WT structure. In detail deformations in the average structure of cluster 3 mainly affect the distal and n-src loops' distances (ABC), while in clusters 4 and 5 the distances of the RT loop from both the other loops and the reference point A (ACD and ABD) depart from the WT geometry.

A closer look at the cluster composition shows the ability of the SOM to group conformations common to all domains, as well as to correctly separate the typical dynamics of each of the three domains with higher flexibility. The contributions of each mutant ensemble to the five clusters (Figure 3.7A) highlights that clusters 1 and 2 are populated by conformations from all the mutants. The larger contributions to cluster 1 are from the WT SH3 and the mutants with reduced conformational flexibility (R21A, N47A and A56G), while cluster 2 is more representative of R21G and N47G ensembles. Each of the remaining three clusters is dominated by one contribution: cluster 3 mainly by A56S, cluster 4 by N47G and cluster 5 is almost completely populated by conformations from the R21G ensemble.

An interesting feature arises from the topological nature of the SOM. Conformational transitions that occur in consecutive times along the MD trajectory involve conformations assigned to neighbour clusters on the map. This can be shown by annotating the clusters detected by the SOM on the RMSD plots. An example for three trajectories is reported in Figure 3.9 where cluster attributions are shown in colour. In the first part of the WT RMSD plot, frequent transitions occur between conformations in the green and blue clusters, that are neighbour in the SOM. More clearly, in the A56S plot, transitions between conformations in the green and yellow clusters, that in the SOM are separated by the violet cluster,

always occur in the trajectory through sampling of conformations in the violet cluster. In the R21G plot, while some transitions occur between conformations in neighbour clusters (blue and green), others (green to yellow or red to yellow) are separated by the violet cluster in the map and occur only through brief sampling of conformations of this type.

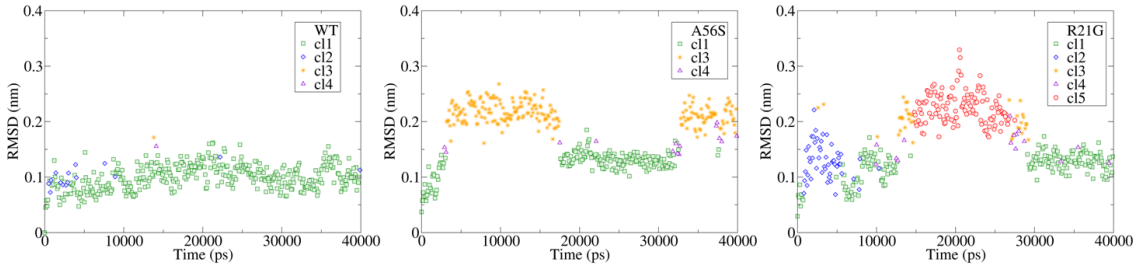



*Figure 3.9 - Plot of RMSD versus time during the MD simulations. From top to bottom, MD trajectories for the WT SH3, A56S and R21G mutants. Conformations attributed to the five clusters obtained from the SOM trained on the entire group of trajectories are colored according to Figure 3.7*

Previous studies suggested an hypothesis on the role of conformational flexibility in reducing the binding affinity (38). To verify this we studied the effects of flexibility on the binding site geometry. As previously described, the binding pocket of the p41 peptide is flanked by the RT and n-src loops (Figure 3.1B), whose dynamics affects the binding specificity. Therefore, the inter-residue distances in the binding site of the SH3:p41 complex (all the heavy atoms in the residues interacting with p41, shown in Figure 3.1B) were compared to the corresponding distances in the representative conformations of each cluster. The dRMSD values confirmed that the increase of conformational freedom induced by the N47G and R21G mutations (mainly described by clusters 4 and 5) produces a significant distorsion of the binding site geometry (dRMSD = 1.7 and 1.8 Å), while other mutations did not produce comparable effects (dRMSD around 1 Å).

### *3.4 Conclusions*

The results of the application of the proposed approach to compare conformational ensembles of protein domains to the SH3 domain and its mutants highlighted the specific advantages of a SOM approach in conformational and functional analysis.

The major benefit is the possibility of providing a topological mapping of the conformational space embedded in a simple 2D visualisation. This simplifies the identification of differences in the conformational dynamics of each domain (see Figures 3.4-3.7). Moreover, the map can adapt to record differences in both large and small fluctuations, as well as to group conformations associated to different directions of the same motion.

The combination of SOMs and complete linkage clustering on the neuron vectors showed good performance in the analysis of single trajectories of the test case. For example, for R21G (Figure 3.4) clusters containing conformations that deviate from the WT equilibrium structure and are associated with loop motions that affect ligand binding were clearly detected. More interestingly, the method resulted in a very efficient comparison of multiple trajectories. In this case, low fluctuations, large concerted motions and intermediate dynamic perturbations were clearly and correctly detected (Figures 3.7 and 3.8, and Table 3.3). The comparison of inter-residue distances in the binding site (75) among the cluster representative conformations led to a functional interpretation of the observed differences. The increase of conformational freedom induced by the N47G and R21G mutations induces a distortion of the binding site geometry that explains the decreased ligand binding ability, while other mutations do not produce comparable effects.

Interestingly, as shown by the annotation of the cluster identity on the RMSD plots of the MD trajectories (Figure 3.9), conformational transitions during the

MD simulations are indirectly recorded on the map: neurons describing conformations involved in a transition are adjacent on the map.

# Chapter 4

## THE ROLE OF FLEXIBILITY IN PROTEIN BINDING: TRANSIENT COMPLEXES OF RAS PROTEINS

### *4.1 Selection of the study case*

As anticipated in Chapter 1, we were interested in the investigation of the role of flexibility in protein binding and in particular its effects on: the bound and unbound forms of transient complexes which show large conformational changes at the interface; the 'promiscuity' in binding in proteins with multiple partners, termed hub proteins

To select complexes that are appropriate study-cases for these purposes, we inspected the Mintsteris databases of transient complexes (79) and the PiSite database (80). In this last one, proteins from the PDB are clustered in families with high sequence identity and the number of binding partners and binding modes is recorded. From this analysis it emerged that the members of the Ras superfamily are good candidates.

The Ras superfamily of small guanosine triphosphatases (GTPases) comprises over 150 human members, with evolutionarily conserved orthologs found in *Drosophila, C. Elegans, S. cerevisiae, S. pombe, Dictyostelium* and plants. Even if a definitive classification of these GTPases is not yet possible, the Ras superfamily has traditionally been divided into five different major branches: Ras sarcoma (Ras), Ras homologous (Rho), Ras-like proteins in brain (Rab), Ras-like nuclear (Ran), ADP-ribosylation factor (Arf). This classification is based on structure, function, or both (81).

To summarize the characteristics of the Ras signaling pathway three main features can be reported. First the Ras family of proteins is ubiquitously expressed in brain,

regulating information processing in many brain regions. They are also interconnected with a large molecular network of upstream and downstream signaling elements. Second, the apparent complexity of this pathway is associated with an excellent organization: a) multi-domain interaction with Guanine nucleotide Exchange Factor (GEFs) and GTPase-Activating Protein (GAPs) is driven by specific signaling messengers and structural proteins, b) clustering of multiple signaling elements by scaffolding proteins. Third, the activity of Ras family proteins is highly dynamic (82).

These evidences allow the classification of this superfamily in the "sociable hubs proteins" (8). The main features of this group of proteins are: a) the ability of interact with multiple partners, changing dynamically the partners; b) a general absence of many disordered regions in the binding interface; c) an high degree of global flexibility.

The basic structure of the GTPase domain in these proteins was first observed in the human protein, H-Ras, and consists of a central six stranded β-sheet (β1-β2-β3 antiparallel, β4-β5-β6 parallel) and five α-helices (Fig. 4.1). The domain is also characterized by five conserved sequence motifs (G1 - G5). Close to G1 and G2, there is a structural region designated as "switch region", including the Switch I, Switch II and P-loop elements (see Fig 4.1). The name "switch" comes from a comparison of the structures of Ras in the GTP- and GDP-bound forms, where it was seen that these are the areas that change most significantly on GTP hydrolysis. Ras proteins always cycle between an active conformation (GTP binding) and an inactive conformation (GDP binding). These two different conformations are shown in Fig. 4.1 for the H-Ras, and in the following they will be named "open" and "close", respectively, with reference to the switch region arrangement. These conformations are also important in defining different binding modes in the formation of transient complexes with other proteins (82).

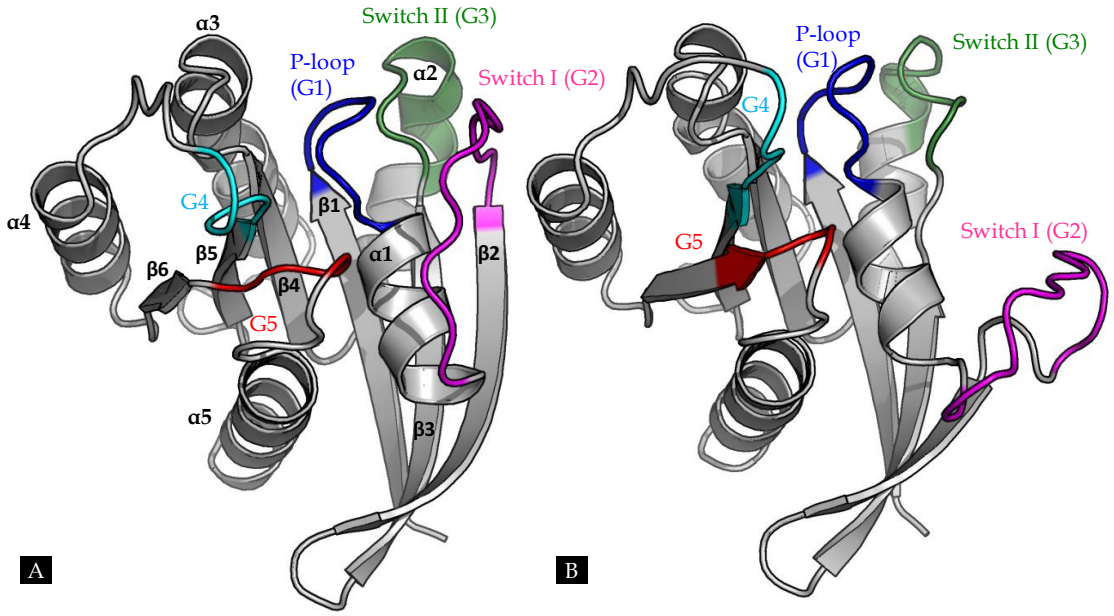



*Figure 4.1: Basic structure of H-Ras: A) in "close" conformation; B) in "open" conformation. Conserved sequence motifs: G1 - often referred as P-loop with the consensus sequence $Gx_4GKS/T$ which is involved in a number of backbone interactions with the nucleotide phosphates; G2 - in the so-called switch I loop; G3 in the so-called switch II region. G4 and G5 are involved in interactions with the guanine base and are responsible for the discrimination against other nucleotides such as ATP/ADP* (82)

Among the Ras complexes identified by the database analysis, four were selected on the basis of the following criteria:

- The crystal structure of the Ras protein involved in the complex is available for both the bound and the unbound forms.
- Large conformational changes are observed between the bound and the unbound forms at the protein-protein interface.

The selected complexes are listed in Table 4.1.

*Table 4.1: Complexes chosen for the analysis: PDB code of the bound and unbound forms of the Ras proteins and of the effectors*; RMSD on Cα for the whole domain and for the residues involved in the interface.*

| Protein complexes | PDB code_chain (Effector) | PDB code_chain (Ras bound) | Ref | PDB code_chain (Ras unbound) | Ref | RMSD Cα (nm) | RMSD interface Cα (nm) |
|---|---|---|---|---|---|---|---|
| Ran / Regulator of chromosome condensation (RCC1) | 1I2M_B | 1I2M_A | (83) | 1QG4_A | (84) | 0.11 | 0.03 |
| Ran / Importin Beta | 1IBR_B | 1IBR_A | (85) | 1QG4_A | (84) | 0.40 | 0.08 |
| H-Ras / Son of sevenless (SOS-1) | 1BKD_S | 1BKD_R | (86) | 1CTQ_A | (87) | 0.32 | 0.19 |
| Rab21 / Rabex-5 catalytic core | 2OT3_A | 2OT3_B | (88) | 1YZU_A | (89) | 0.31 | 0.17 |

**The missing residues in the PDB structures were modeled by using Modeller 9v7* (90)

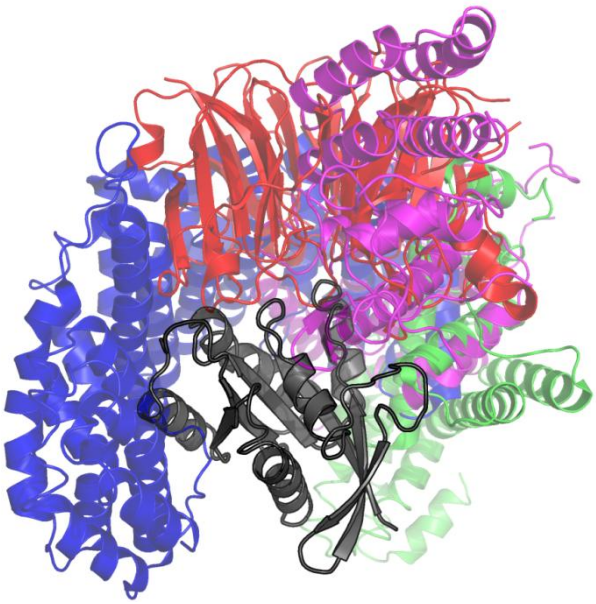


*Figure 4.2: Multiple structural superimposition of the selected complexes. The structure of H-Ras (1BKD_R) is shown as reference for all the complexes. The effectors are: Son of Sevenless (SOS-1) in magenta (1BKD_S), Regulator of chromosome condensation (RCC1) in red (1I2M_ B), Importin beta in blue (1IBR_B), Rabex-5 catalytic core in green (2OT3_ A).*


*Figure 4.3: Multiple sequence alignment of the selected cases.*

As expected from the "sociable hub" properties of Ras proteins, an interesting characteristic of these complexes is that different regions in the Ras surface have the role of interface in the different complexes (Fig 4.2).

Comparing the members of this group of Ras domains, it can be observed that the sequence is not so well conserved (Fig 4.3) (sequence identity c.a. 20%) but, as expected, the general fold is more conserved.

A structural comparison between the bound and the unbound structures was performed. Moreover, a structural analysis of the interfaces was performed with the parameter optimized surface (POPS) method (91) by calculating the solvent accessible surface area (SASA) that is buried upon complex formation.

The overall conformational change of each system is summarized in Table 4.1 by the RMSD value on the Cα for the whole domain, and by the same parameter for the residues involved in the interface. A more detailed comparison is illustrated by Fig 4.4. In the upper part, the plot of the RMSD on the Cα atoms between the bound and the unbound structures versus residue position is shown. The residues involved in the interface are highlighted in the plot. In the lower part of the Figure, the superimposition of the bound and unbound structures is shown, with interface residues highlighted.

This analysis indicates that these Ras proteins undergo a large conformational change upon binding in all the selected complexes. Referring to the numbering attributed after sequence alignment (Fig 4.3), the switch region lies at the N-term in the range of residues 9 - 74. In particular, the P-loop includes residues 9-16, the Switch I residues 25-37, and the Switch II residues 57-74. It is evident from the RMSD plots in Fig 4.4 that the largest conformational changes observed (RMSD to 9 Å) are in the switch region. In two cases (H-Ras and Rab) this region adopts the "open" state in the bound form (grey in the bottom part of Fig 4.4) and the "close" state in the unbound one (cyan). For one of the Ran proteins (1IBR_A) the opposite behaviour is observed, while for the other Ran (1I2M_A) both the bound and unbound forms are in the "open" state. As previously observed in Fig 4.2, where

the 3D structure of the four complexes are shown, in the H-Ras/SOS-1 (magenta), Rab21/Rebex-5 (green) and Ran/Importin β complexes the protein-protein interface is located in the switch region or includes a large part of it, while in the Ran/RCC1 complex the switch region is not directly involved in the binding interface.

Following the outcomes of the application of our SOM protocol to the previous case described (Chapter 3), it appears interesting to analyse cases in which relevant conformational changes occur during the simulations. In the selected systems, large conformational differences are observed comparing the bound to the unbound forms. If the MD simulations will be able to sample these changes, it is expected that both the global flexibility of the domains and the local flexibility at the interface will be highlighted by the SOM analysis, thus allowing the detection of interesting features connected to the binding to the various effectors.

The employment of the SOM analysis for these study cases, also involves some interesting methodological aspects. The same protocol that was optimized by using the SH3 study case (Par. 2.3) was employed also in the analysis of these domains. This allowed to test the applicability of the proposed protocol to systems with higher dimensions and with different dynamical characteristics.

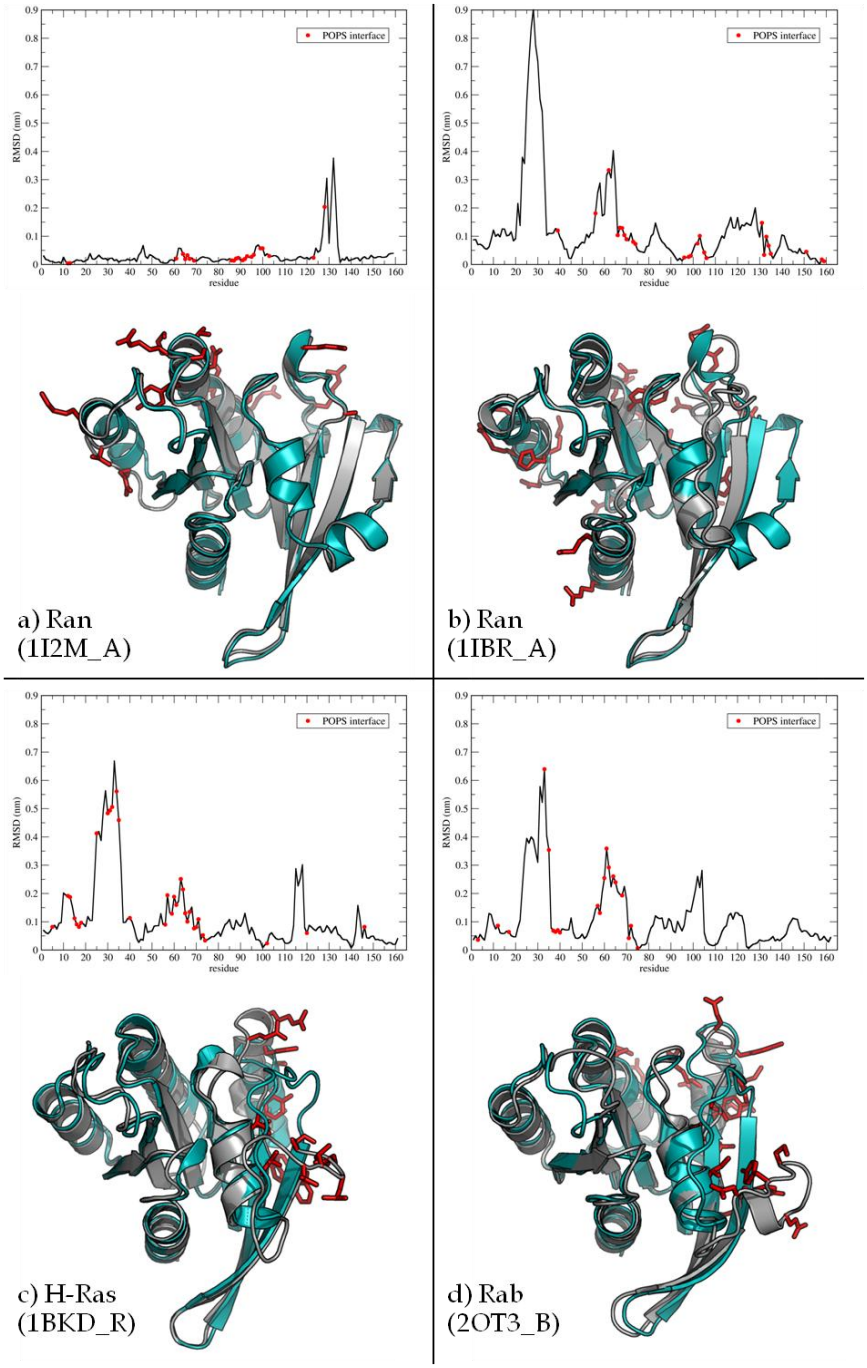



*Figure 4.4: Structural comparison between the bound and the unbound forms of the four Ras proteins. For each protein; Upper part: graph of RMSD on Ca vs. residue position (residues involved in the interface are highlighted with red dots). Lower part: superimposed structures of the bound (grey) and unbound (azure) forms (for the bound structures the residues involved in the interface are shown as red sticks)*

## 4.2 MD simulations of the Ras proteins

### 4.2.1 Simulation protocol

For each complex four different MD simulations were performed (Fig. 4.5):

- Simulation of the complex, starting from its crystal structure ("bound");
- Two separated simulations of the partners of the complex, starting from the single structures of the two partners extracted from the complex ("separated");
- Simulation of the unbound domain of the Ras superfamily member, starting from the crystal structure of the not complexed domain ("unbound").

In the following, each simulation will be labelled with the PDB code_chain (see Tab. 4.1) and a lower-case letter indicating the state: b= bound, s= separated, u=unbound (see the example in Fig. 4.5)

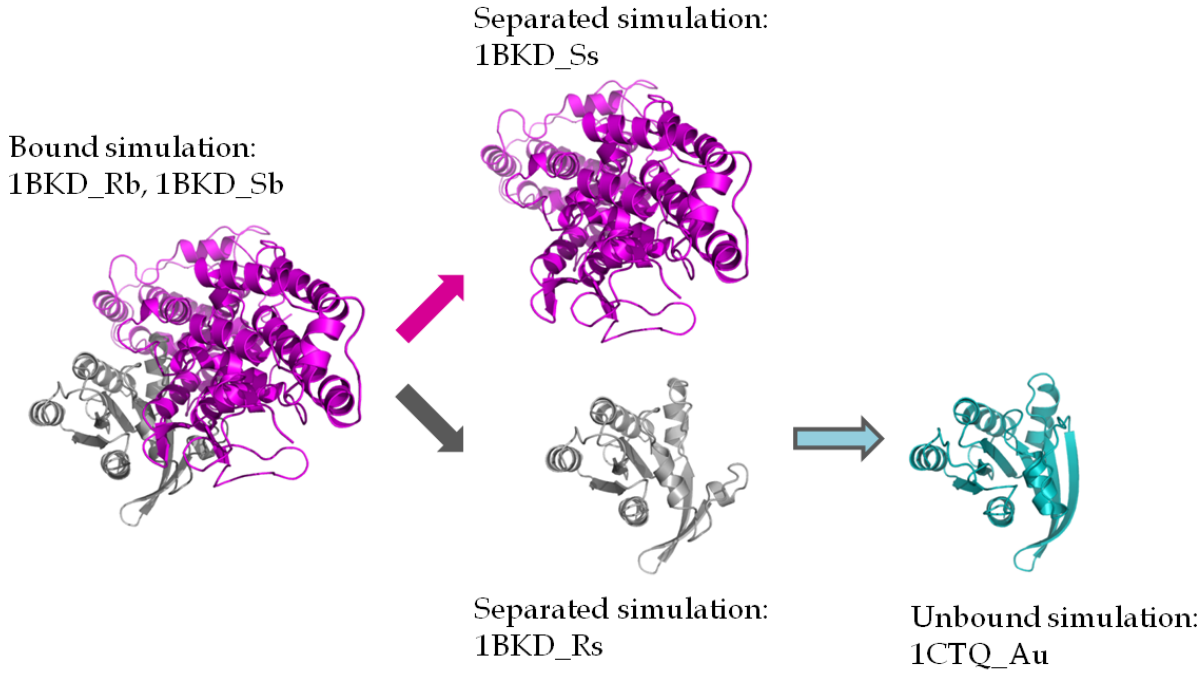



*Figure 4.5: Schematic representation of the simulations performed for each complex. 1BKD complex is used as example.*

Each simulation of a Ras bound form will be compared a) to that of the unbound one, with the aim to analyse the large conformational changes involved in the

binding process; b) to that of the separated form, to analyse the changes in flexibility specifically related to the association/dissociation to the partner, while keeping the common initial structure. As the main focus of this discussion are the changes in flexibility of the Ras proteins, associated to binding, the simulations of the effectors will be briefly presented at the end of the Chapter (Par. 4.4).

The conformational dynamics of the selected systems was analysed by 25 to 75 ns MD simulations.

All the simulations and the subsequent analysis were performed with GROMACS 4.0.1 (44), using the GROMOS96 43a2 force field. The general protocol for system preparation and simulation described in Par. 2.1.2 and 2.1.3 was applied, by setting some parameters specifically for these systems.
Each structure was immersed in a cubic box with SPC water molecules (49) and the box size was set to ensure a distance of at least 1.2 nm between the protein and the box boundaries. Counterions were added to balance the charge of the systems.
The systems were simulated in the NPT ensemble (constant number of atoms, pressure P, and temperature T), and the Berendsen algorithm was employed for temperature (T = 300 K) and pressure (p = 1 bar) regulation, with coupling constants of 0.2 and 1 ps, respectively. Periodic boundary conditions were imposed during the simulations. The equations of motion were integrated using the leap-frog method with a 2-fs timestep. All the bonds in the protein were frozen with the LINCS method (52), while SETTLE (50) was used for water molecules. Long range electrostatic interactions were calculated with the Particle Mesh Ewald (PME) method (40). A 14-Å cutoff was used for all the non-bonded interactions.
Before starting the simulation, the system was minimized with 1000 steps of steepest descent. Then harmonic positional restraints (with a force constant of 4.8 $kcal \cdot mol^{-1} \cdot Å^{-2}$) were then imposed onto the protein heavy atoms and gradually turned off in 180 ps, while the temperature was increased from 200 to 300 K. Finally the system was equilibrated for 2 ns without restraints.

### *4.2.2 Analysis of the trajectories*

The plots of the RMSD values from the starting structure of the simulation, computed on the Cα, versus the simulation time are reported in Fig. 4.6.

No relevant global conformational changes and limited average deviation (between 0.2 and 0.3 nm) from the starting structure are observed for all the systems.

For the first Ran protein (1I2M_A), the three simulations show an average RMSD of 0.2 nm. For both the bound and the unbound simulations the first 2 ns were needed to complete the equilibration of the system. The other Ran system (1IBR_A) shows: a) an higher average RMSD (0.3 nm) in the bound and the separated simulations than in the unbound one, b) the need of 2 ns to complete the equilibrations, c) a possible lack of convergence of the sampling in the separated simulation.

Also for H-Ras both the bound and separated simulations have higher average RMSD compared to that of the unbound one that, after 4ns of equilibration, shows an average RMSD of 0.2 nm. Again in the separated simulation the plot is drifting toward higher values, indicating a possible lack of convergence of the sampling.

For the Rab protein, the three simulations show similar average values and tendencies and all of them show a slight drifting toward higher values of RMSD.

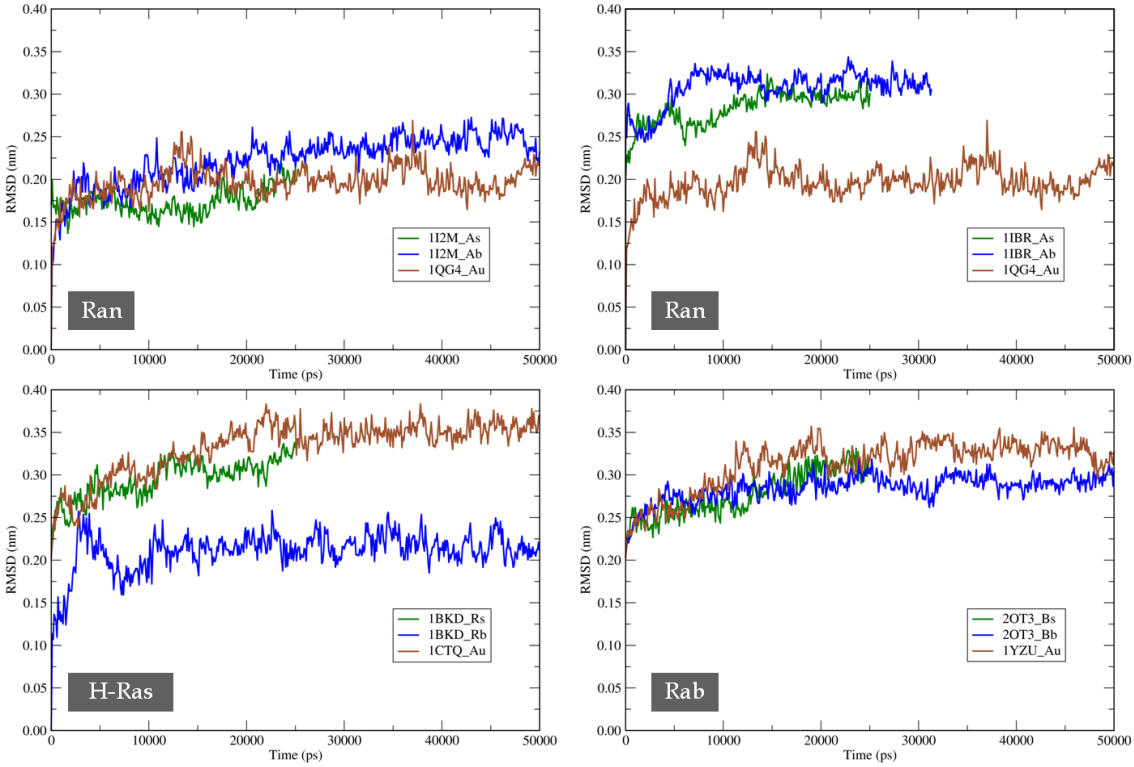



*Figure 4.6: RMSD computed on Ca. For each case, in green the "separated" simulation, in blue the "bound" simulation, and in brown the "unbound" simulation. For each trajectory, the structure used as reference to compute the RMSD values is the conformation of the system at t=0 of the simulation. (to make the comparison easier, in the RMSD plot of 1CTQ_Au the first 50ns of the 75ns of simulation are shown)*

A test on the completeness of conformational sampling was performed by using the overlap between the conformational space spanned in each half of the simulation and that of the overall simulation as a convergence index, as described in Par. 2.1.4. The results, reported in Table 4.2, indicate different levels of overlap with the whole trajectory in the two halves of the simulation for all systems. In particular, in four cases (1I2M_Ab, 1IBR_As, 1BKD_Rs and 1CTQ_Au) there's a high difference between the overlap values evaluated in the two halves: the overlap of the first half is always higher than 60% (values between 0.64 and 0.71), while that of the second half is lower than 60% (0.48 - 0.59). In two of these cases (1BKD_Rs and 1IBR_As) a possible reason of this incomplete sampling could be the shorter length of the simulations (25 ns) compared to the other cases (50 to 75 ns). For these two cases also the analysis of

the RMSD (Fig. 4.6) indicated a not complete convergence of the trajectories. In all the other cases the overlap is higher than 60% (0.60 - 0.76) in both the halves of the simulation.

*Table 4.2: Overlap of sampling in the MD simulations of the Ras domains: overlap between the conformational space spanned by each half of the Ras simulation and that of the overall trajectory. In the last column the length of the simulation.*

| Domain | Case | 1st half | 2nd half | Time (ns) |
|---|---|---|---|---|
| Ran | 1I2M_Ab | 0.71 | 0.55 | 50 |
| | 1I2M_As | 0.62 | 0.65 | 25 |
| | 1IBR_Ab | 0.60 | 0.60 | 50 |
| | 1IBR_As | 0.70 | 0.48 | 25 |
| | 1QG4_Au | 0.76 | 0.67 | 50 |
| H-Ras | 1BKD_Rb | 0.67 | 0.68 | 50 |
| | 1BKD_Rs | 0.68 | 0.59 | 25 |
| | 1CTQ_Au | 0.64 | 0.53 | 75 |
| Rab | 2OT3_Bb | 0.65 | 0.62 | 50 |
| | 2OT3_Bs | 0.62 | 0.64 | 25 |
| | 1YZU_Au | 0.63 | 0.61 | 50 |

*Table 4.3: Ras Essential space: Percentage of total space described using increasing number of eigenvectors*

| | Ran | | | | | H-Ras | | | Rab | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| PC | 1I2M_Ab | 1I2M_As | 1IBR_Ab | 1IBR_As | 1QG4_Au | 1BKD_Rb | 1BKD_Rs | 1CTQ_Au | 2OT3_Bb | 2OT3_Bs | 1YZU_Au |
| 1 | 29.5 | 26.9 | 28.1 | 37.7 | 20.6 | 21.3 | 31.3 | 29.2 | 26.3 | 28.6 | 30.2 |
| 1-2 | 42.6 | 36.0 | 49.1 | 45.4 | 33.3 | 34.8 | 44.8 | 46.0 | 42.2 | 40.4 | 47.6 |
| 1-3 | 49.6 | 43.6 | 55.1 | 51.7 | 39.9 | 41.6 | 52.5 | 53.1 | 48.0 | 48.6 | 52.8 |
| 1-5 | 58.0 | 53.8 | 62.8 | 60.1 | 48.8 | 51.0 | 61.3 | 63.0 | 56.8 | 57.3 | 60.9 |
| 1-10 | 68.3 | 65.3 | 71.5 | 70.6 | 62.4 | 63.6 | 73.6 | 73.6 | 68.3 | 69.1 | 71.6 |
| **1-15** | **74.6** | **72.0** | **76.8** | **76.3** | **70.4** | **71.1** | **79.5** | **79.4** | **74.5** | **75.3** | **77.3** |
| 1-20 | 78.6 | 76.4 | 80.2 | 80.1 | 75.3 | 75.8 | 83.2 | 83.2 | 78.6 | 79.4 | 81.1 |
| 1-50 | 89.1 | 87.9 | 90.0 | 89.9 | 87.9 | 88.2 | 92.1 | 92.6 | 89.7 | 89.9 | 90.9 |
| 1-all | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |

The most informative directions of motion were extracted by Essential Dynamics analysis (see Par. 2.1.5) and the fraction of total motion described by different subspaces was evaluated to identify the extent of the essential space for each domain. Results reported in Table 4.3 indicate that in all cases more than 70% of the motion is described by the first 15 eigenvectors (corresponding to about the 3% of the total space). Therefore a 15 dimensional essential space was selected for the following analyses.

### *4.3 Conformational and functional analysis of Ras proteins by SOM clustering*

In this paragraph the role of flexibility in the binding of the four selected Ras proteins will be described, first by using the traditional techniques for MD simulation analysis (described in Par. 2.1.5 and 2.1.6): the RMSF on the positions of the Cα atoms, calculated in the essential space; and the analysis of the first principal component derived from the ED analysis. For each Ras protein, the three simulations described in Fig. 4.5 will be discussed and compared.

Then, the SOM approach developed in this thesis (see Par. 2.3) will be applied: for each case, the clustering of both single trajectories and multiple trajectories of the same domain, in the different conditions of simulation ("bound", "separated", "unbound"), will be presented. From a methodological point of view, the SOM analysis of these trajectories allowed us to answer two questions about the exportability of the protocol to domains different from the case used for its development: 1) a question regarding the SOM learning, i.e. are the parameters designed for a domain of 55 residues appropriate also in the analyses of domains three times bigger? 2) a more general question about the representation chosen: are the Cα coordinates a good descriptor also in these cases?

The results for the four Ras domains will be grouped according to some characteristics of the binding interface that greatly influence the dynamic behavior. As described in Par. 4.2, the Ran (1I2M_A) domain differs from the other domains, as the switch region is only partially involved in the binding interface with RCC1 and, as a consequence, it maintains the same "open" arrangement both in the bound and in the unbound forms (Fig 4.4a). The role of its typical flexibility in binding will be discussed in Par. 4.3.1. On the contrary, the other three proteins, Ran (1IBR_A), H-Ras and Rab (Fig 4.4b-d), constitute a group where the switch region is in part (Ran(1IBR_A)) or entirely (H-Ras and Rab) involved in the binding interface with the effectors. Therefore, this region adopts different

arrangements in the bound and the unbound forms of these domains. The results will be summarized in Par. 4.3.2.

### *4.3.1. Ran (1I2M_A) dynamics*

The trajectories used in this case are: 1I2M_Ab, 1I2M_As and 1QG4_Au.

The first analysis includes the comparison of the local flexibility per residue through the RMSF plots of the three simulations (Fig. 4.7), and the analysis of the most relevant collective motions, as described by the first principal component (PC1) (Fig. 4.8). For this domain, the PC1 describes at least the 20% of the overall motions in all the three simulations (Tab 4.3).

The flexible regions in the 1I2M_A simulation (fluctuations higher than 0.1 nm) include some connecting loops (around residue 50, 110, and 145) along with four regions that are the most interesting for the comparison with the other simulations: p-loop, switch I, switch II and residues in the range 120-140. To make the interpretation of the differences in flexibility easier, these regions are highlighted both in Fig 4.7 and in Fig 4.8 by using the same color code in all the representations: p-loop in blue, switch I in magenta, switch II in green, residue 120-140 in cyan.

The comparison between the RMSF plots of the bound and unbound simulations (blue and brown profiles in Fig. 4.7) highlights different dynamical behaviours associated to binding, but also reflects the differences in the initial X-ray structures. In this case, structural differences (of about 0.4 nm) were observed only in the region including the small helix around the position 130 (see Fig. 4.4a), that in the bound structure (1I2M_A) is more disordered than in the unbound one (1QG4_A). As a consequence, the flexibility of the region including residues 120-140 is higher for the bound system than for the unbound, with the exception of a higher peak near the interface residues in the unbound simulation. For the rest of the domain, the unbound simulation profile presents more pronounced peaks than the bound one in the regions including interface residues (p-loop, switch II and the long

central helix). This behaviour can be associated to the effects of the lack of the binding partner on the domain flexibility.

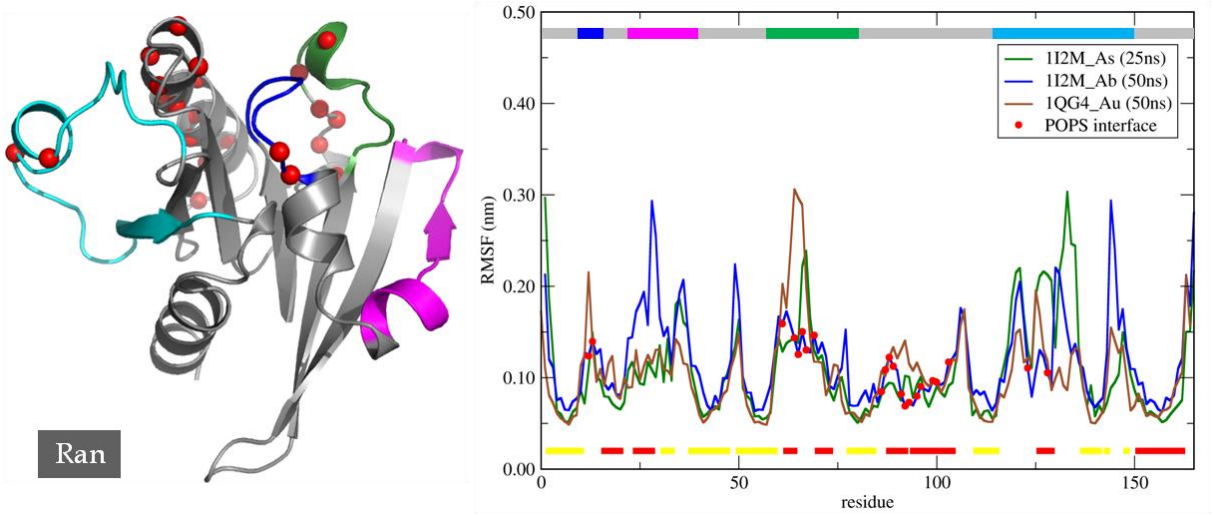



*Fig 4.7: Ran (1I2M_A) RMSF analysis. On the left, the 3D structure of the bound domain with the most interesting flexible regions highlighted (blue = p-loop, magenta = switch I, green = switch II, cyan = residues 120-140). On the right, the RMSF profiles of the three simulations ("bound" in blue, "separated" in green, "unbound" in brown); at bottom, the secondary structure of the "bound" conformation at t=0 of the simulation, attributed by the DSSP program* (92) *(red = helixes, yellow = sheets); at top, bar showing the most interesting flexible regions, colored using the same color-code of the 3D structure. Both in the 3D structure and in the RMSF plot the residues involved in the interface are highlighted with red dots.*

The comparison of the RMSF plots of the bound and the separated simulations (blue and green profiles in Fig. 4.7) allows to highlight the differences in flexibility caused by the removal of the partner protein, while maintaining the same starting structure. It can be observed that in two regions that are partially involved in the binding interface, the C-term part of the switch II and residues in the range 125-140, the domain flexibility is higher in the separated than in the bound form. On the contrary, the switch I region, that doesn't include interface residues, results much more flexible in the bound simulation and in two loops (around position 50 and 145) that are far from the bound region.

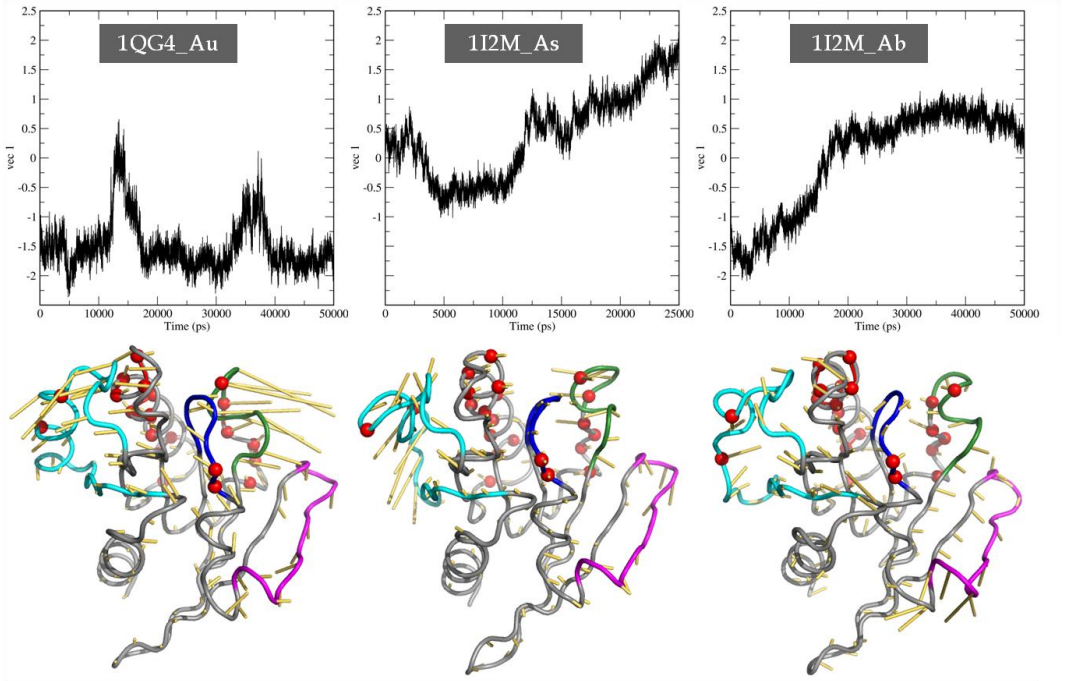



*Fig 4.8: Ran (1I2M_A) PC1 analysis of the three simulations. In the upper part, the projections of the first PC during the simulations. In the lower part, the projection of the motion described by PC1 on the 3D structure. The average conformation during the simulation is represented as tube, with the most interesting flexible regions highlighted (blue = p-loop, magenta = switch I, green = switch II, cyan = residues 120-140); the residues involved in the interface are shown as red dots. The yellow line associated to some residues, describes the direction and the amplitude of the displacement described by PC1 during the trajectory for that residue.*

The principal motions described by the PC1 (Fig. 4.8) are different for the unbound state with respect to the other simulations. For 1QG4_Au the PC1 describes the high amplitude concerted motion of the switch II and the 120-140 regions (as described by the length of the yellow needles in Fig. 4.8). For 1I2M_As it mainly represents a motion of the 120-140 region that gradually moves away from the binding zone. In 1I2M_Ab the PC1 describes an overall slight fluctuation of the whole domain, constrained by the presence of the binding partner, and an higher amplitude motion in the switch I.

Once described the flexibility of the domain using both the RMSF profiles and the PC1 analysis, the simulations were studied using the SOM approach. For this purpose, following the protocol presented in Par. 2.3, each sampled conformation was described by the Cartesian coordinates of the 162 Cα atoms corresponding to

structurally equivalent positions in all the domains (see the alignment in Fig. 4.3). Therefore, the input data presented to the SOM were vectors of 486 elements. As shown in Table 4.2, these simulations have different length, from 25 to 50 ns. Using the sampling rate of 1/100 ps (Par. 2.3.5) a different number of input data vectors is obtained for each simulation (250 vectors for 1I2M_As, 500 vectors for 1I2M_Ab and 1QG4_Au).

First we analyzed the single trajectories to describe the conformations that best summarize the conformational sampling of each simulation.

By applying the Mojena's rule (69) after hierarchical clustering of the obtained neurons, all the maps of the single trajectories were divided in four clusters (Fig. 4.9).

For the unbound system (1QG4_Au), the previous analyses suggested that the region with the highest fluctuation is the switch II. In fact, the comparison of the conformations of the centroids of the clusters in the obtained map (the upper part of Fig. 4.9), with respect to the reference structure, indicates that clusters 1 and 2 describe the progressive motion of this region toward the switch I. In both the centroids (red and brown) it is possible to see that the two switches are closer one to the other compared with the reference structure (black). The other two clusters, clusters 3 and 4 (in the upper part of Fig 4.9), in which the switch region has the same conformation of the reference structure, mainly describe the different conformations of the region between residues 120 and 140 on the opposite side of the domain.

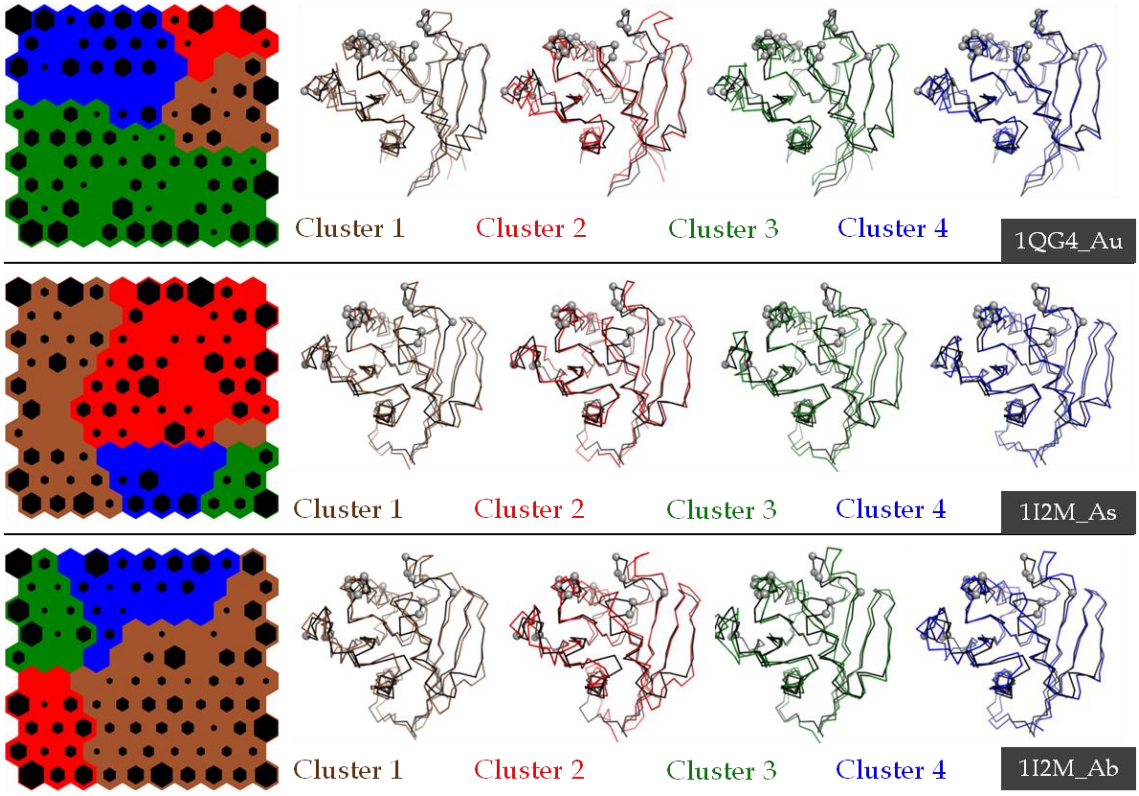



*Figure 4.9: Ran (1I2M_A) SOM analysis – In each section: on the left, the representation of the bidimensional output SOM: the black hexagon in each neuron represents its number of hits, the clusters obtained by hierarchical clustering of the neurons are indicated by different colors. On the right, the ribbon representation of the conformation of the centroid of each cluster superimposed to a reference structure (conformation at t=0 of the MD simulation), reported in black. The residues involved in the interface are highlighted by grey spheres.*

In the SOM analysis of the separated simulation (1I2M_As) (in the middle part of Fig. 4.9), the different conformations of the switch II sampled during the trajectory are described by cluster 2. The other three clusters are needed to describe the different arrangements of the domain in the region between the residues 120 and 140. Interestingly the p-loop conformation is different to the reference conformation in the centroids of all the clusters.

From the SOM analysis of the bound simulation (1I2M_Ab) (in the lower part of Fig. 4.9) it emerges that, within the region involved in the interface, both the switch II and the p-loop have conformations different to the reference structure in all the clusters. The motions of these two elements seem to be correlated, in fact their

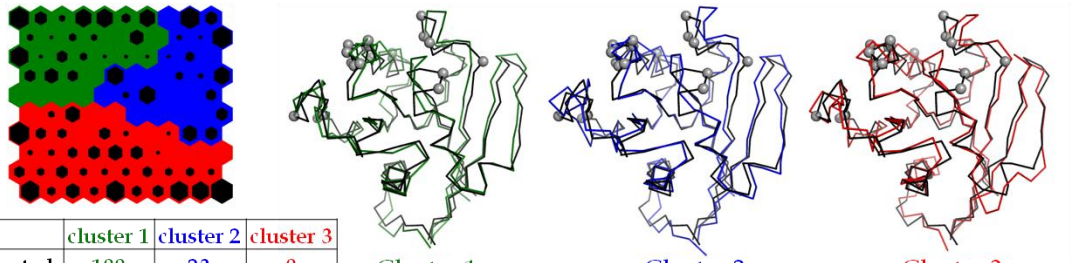
relative conformations in the four centroids describe a progressive separation, from cluster 1 to 4. Moreover, also the conformations of the region between residues 120 and 140 is an important driving force in the conformational learning process of the SOM (see clusters 3 and 4). Another relevant source of conformational changes is the fluctuation of the N-term of the switch I (see clusters 1 and 2 in the lower part of Fig 4.9).

The next step was the analysis of multiple trajectories, to detect similarities and differences in the conformational sampling. A first comparison between the conformations sampled during the unbound and the others simulations suggested us that the conformational basin sampled during the unbound trajectory is not directly comparable with those of the separated and the bound simulations. Remembering that the starting structure of the unbound system (1QG4_A) was different from the bound one (Fig. 4.4a), in particular around position 130, this difference seems to be maintained during the simulations.

This condition is not present in the comparison between the separated and the bound simulations, that both start from a common structure (1I2M_A).

The dimension of the matrix given in input to the SOM for analysing the combined trajectories was of 750 vectors (each composed by 486 elements), 250 of the conformations extracted from the separated simulation plus 500 of those extracted from the bound simulation.

The cluster analysis of the obtained map (Fig. 4.10) indicates three clusters and, interestingly, one of them (cluster 2) is populated by conformations belonging to both the bound and the separated simulations. The analysis of the other two clusters indicates that the presence/absence of the partner (RCC1) seems to produce the sampling of different conformational regions in the two trajectories.




| | cluster 1 | cluster 2 | cluster 3 |
|---|---|---|---|
| separated | 100 | 23 | 0 |
| bound | 0 | 77 | 100 |

*Figure 4.10: Ran (1I2M_A) SOM analysis: comparison between separated and bound simulations. On the left, the representation of the bidimensional output SOM (see Fig. 4.9 for the details) and, at bottom, the percentage distribution of conformations of each simulation in the clusters. On the right, the ribbon representation of the conformation of the centroid of each cluster, superimposed with a reference structure (conformation of 1I2M_As at t=0 of the MD simulation), reported in black. The residues involved in the interface are highlighted as grey spheres.*

In particular, the effect of the absence of the partner is described by the conformations in cluster 1, that is populated only by conformations of the separated simulation. On the contrary, the effect of the presence of the partner is described by cluster 3, in which there are only conformations of the bound simulations. Analysing the conformation of each centroid, cluster 1 slightly differs from the reference structure only for the conformations of switch II and p-loop, therefore describing the fluctuations around the equilibrium position of the separated simulation (Fig. 4.9). In cluster 2 the differences in those two region are more evident, and there is also a different arrangement of the region between residues 120 and 140, whose flexibility is characteristic of both the bound and separated simulations. Cluster 3, that describes the typical conformations sampled in the bound simulation, includes, along with different conformations of p-loop and switch II, a different local arrangement of switch I at its N-term. This cluster summarizes the fluctuations observed for this domain in the SOM analysis of the single trajectory (lower part of Fig. 4.9).

Therefore, from the analysis of the combined trajectories the characteristic features of the two domains emerge in cluster 1 and 3, i.e. the switch II – p-loop motion for the separated simulation and the flexibility at the N-term of switch I for the bound

one. Moreover a common conformational basin is identified and described by the cluster 2.

### *4.3.2. Ran (1IBR_A), H-Ras and Rab dynamics*

The trajectories used in this case are: 1IBR_Ab, 1IBR_As and 1QG4_Au for the Ran domain, 1BKD_Rb, 1BKD_Rs and 1CTQ_Au for H-Ras and 2OT3_Bb, 2OT3_Bs, 1YZY_Au for Rab.

As described above, these domains differ from the previous case (Ran – 1I2M_A) for the difference between the bound and the unbound form, especially in the switch region (Fig 4.4b-d).

Also for this group of domains the first analysis is the comparison of the three simulations ("bound", "separated", "unbound") of each domain, to describe the local flexibility on a residue base. Instead the analysis of the first principal component (PC1) and the SOM analyses of the single trajectories is reported only for H-Ras, taken as an example of the behavior of the motions for this group of cases.

The results will be discussed following this schema: first, the RMSF comparison between the bound and unbound simulations of each case, to discuss the different flexibility due to both the binding and the different crystal structure; second, the RMSF comparison between bound and separated simulations, to highlight the differences in flexibility coupled with the presence/absence of the effectors; third, the PC1 analysis of H-Ras to describe its principal motion (this PC allows to describe at least the 21% of the overall information, see Table 4.3); fourth, the SOM analyses of the single trajectories of H-Ras; and finally the SOM analyses of the combined trajectories of the bound and separated simulations of all the cases.

For all the cases in Fig. 4.11 and 4.12, to make the interpretation of the differences easier, the same color code is used to describe the most interesting flexible regions: p-loop in blue, switch I in magenta, switch II in green, residues 102-105 in orange

and regions 120-130 in cyan (this last region in Ran - 1IBR_A includes more residues, 120-140).

From the analysis of the RMSF in the essential space (Fig. 4.11) it emerges that for all the systems the highest fluctuations in the switch regions reach about 0.30 - 0.35 nm, thus indicating that in all the simulations the systems do not sample the conformational change required to pass from the open to the close conformation or *vice versa.* In fact, the expected amplitude of the motion from the bound to the unbound forms of the starting crystal structures is between 0.7 and 0.9 nm, as shown in Fig 4.4. On the contrary, the switches fluctuate around their starting conformations.

The flexible regions of Ran (1IBR_A) include, in addiction to the most interesting regions highlighted by colors: the connecting loop between switch I and switch II around position 50, the long helix around position 80 and the following loop around position 100; all of them include some interface residues.
In the comparison between the bound and the unbound simulations, we have to consider also the relevant conformational differences between the two initial crystal structures (see Fig. 4.4b), that mainly involve the switch I (0.9 nm) and the switch II (0.2 - 0.3 nm).
Despite these clear differences, the overall RMSF profile, in terms of location of the peaks, is conserved (brown and blue plots in Fig 4.11). However, locally the amplitudes of some peaks are different.
The relative flexibilities of the p-loop and the switch I are inverted: the p-loop is more flexible in the unbound simulation, the switch I in the bound simulation. In fact in the starting crystal structure of the bound Ran (1IBR_A) the switch I is in "close" conformation, faced to the p-loop, whereas in its unbound structure (1QG4_A) the switch I is "open" and there are no interactions with the p-loop. Also the switch II is more flexible in the unbound simulation. This region is involved in the interface of the crystal structure of the complex Ran/Importin β

,thus reducing its freedom, and is more disordered in unbound form of Ran (1QG4_A), thus increasing its freedom (Fig 4.4b).
The comparison of the separated (green) and bound (blue) profiles of Ran (upper part of Fig 4.11) indicates that, as expected, the differences are located in regions involved in the interface with Importin β. In fact, the flexibility between residue 120 and 140 is clearly higher in the separated simulation. More slight differences involve the long helix around position 100.

In the Rab RMSF profile (in the middle part of Fig 4.11), the regions listed as relevant for this group of domains are confirmed. In the comparison between the bound and unbound forms, the regions with the highest displacement between the two crystal structures (Fig. 4.4d) are the switch I (0.60 nm), the switch II (0.30 nm) and the helix around position 100. The interface with Rabex5 involves both the switches and the p-loop regions.
Differences in these regions are also evident in the RMSF profile (middle part of Fig. 4.11) with different relative flexibilities (brown and blue plots). As expected, the flexibility is higher in the unbound simulation in the switch I region and in the loop around position 105 (orange region). Interestingly the bound simulation shows an higher flexibility in the switch II, that is totally involved in the interface.

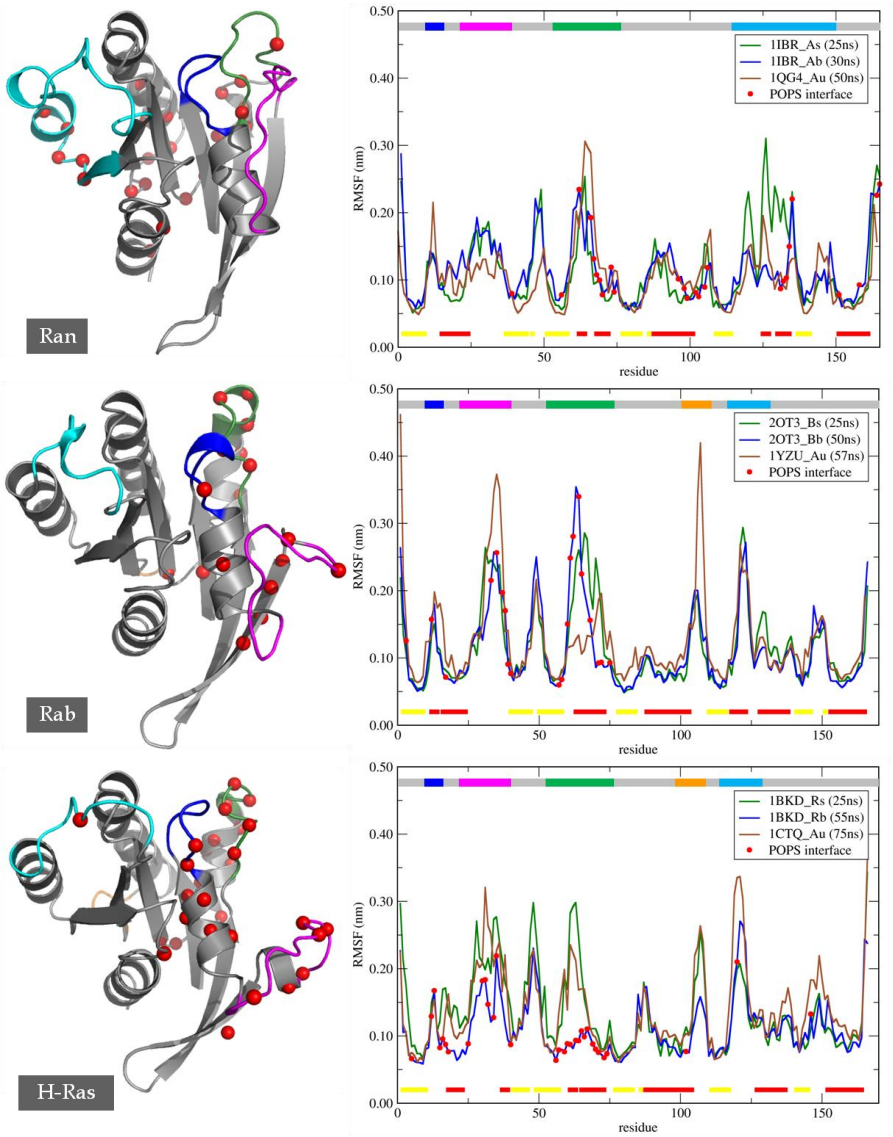



*Fig 4.11: Ran (1IBR_A), Rab and H-Ras RMSF analysis. For each domain: on the left, the 3D structure of the bound domain, with the most interesting flexible regions highlighted (blue = p-loop, magenta = switch I, green = switch II, cyan = residues 120-130, orange = residues 102 -105) On the right, the relative RMSF profiles of the three simulations ("bound" in blue, "separated" in green, "unbound" in brown); at bottom, the secondary structures of the "bound" conformation at t=0 of the simulation, attributed by the DSSP program* (92) *(red = helixes, yellow = sheets); at top, a bar showing the most interesting flexible regions, colored using the same color-code of the 3D structure. Both in the 3D structure and in the RMSF plot the residues involved in the interface are highlighted with red dots*

The comparison between the bound (blue line) and the separated (green line) indicates that the regions with the most relevant differences are the switches. In the switch I region even if the amplitude of the highest peak is the same (0.25 nm) the shape of the profile seems is different. Unexpectedly, in the switch II that is completely involved in the interface with Rebex-5, the RMSF profile is not conserved and the maximum fluctuation has highest values in the bound simulation.

The main flexibility of H-Ras (lower part of Fig. 4.11) is in the regions listed above and also in some connecting loops (around residues 50, 80 and 150). The interface of the complex with SOS-1 directly includes the switches and the p-loop. In these regions the crystal structures of the bound and the unbound forms (Fig 4.4c) are consequently different, with displacements of c.a. 0.20 nm in the p-loop and switch II, 0.60 nm in the switch I, 0.30 nm in the loop around position 120.
In the RMSF comparison (lower part Fig 4.11) between the bound (blue) and the unbound (brown) plots, a general higher flexibility of the domain is observed in the unbound simulation, including all the interface regions.
The comparison between the RMSF profiles (lower part of Fig. 4.11) of the separated (green) and bound (blue) forms indicates that the removal of the partner (SOS-1) causes an higher flexibility. Focusing the attention on the region involved in the interface, in the switch II the difference is higher, with a peak of 0.30 nm in the separated simulation. The only exception, in which the bound simulation has higher flexibility compared with the separated is around position 105 (cyan region).

For H-Ras also the PC1 analysis is reported (Fig 4.12). As in the case of Ran (1I2M_A) the first principal components of the three simulations explain at least the 20% of the overall flexibility. The principal motions described by PC1 in the unbound are different to the other simulations.

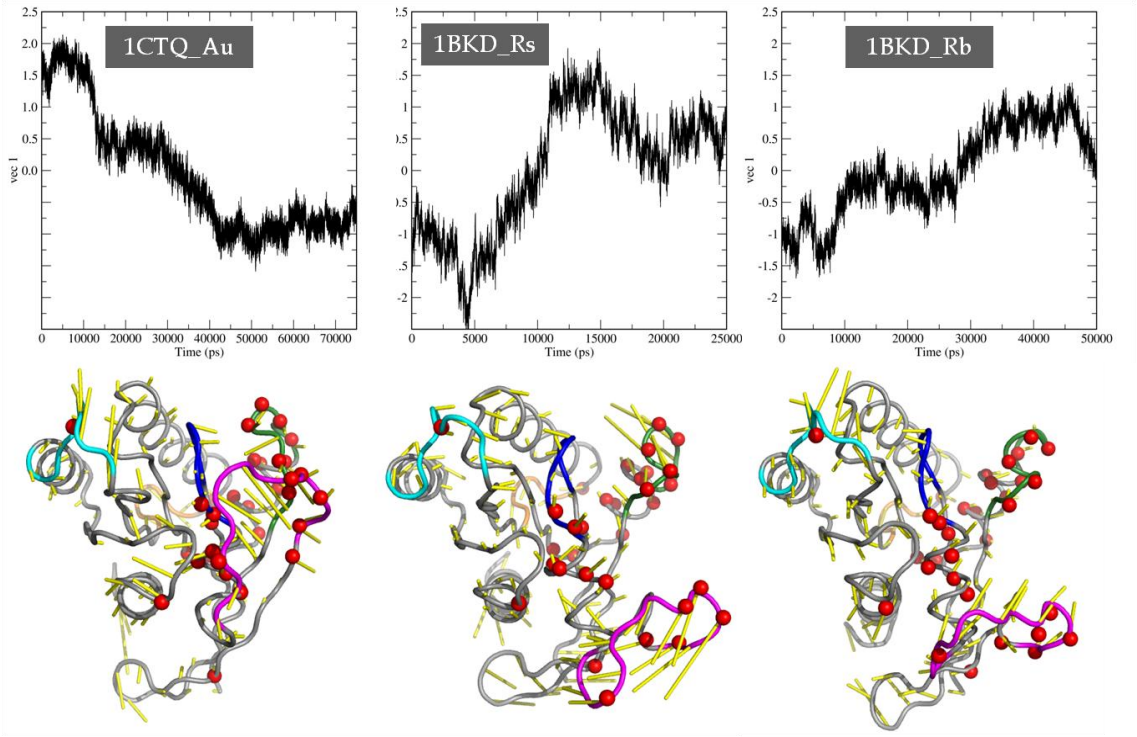



*Fig 4.12: H-Ras PC1 analysis of the three simulations. In the upper part, the projections of the first PC during the simulations. In the lower part, the projection of the motion described by PC1 on the 3D structure. The average conformation during the simulation is represented as tube, with the most interesting flexible regions highlighted (blue = p-loop, magenta = switch I, green = switch II, cyan = residues 120-140, orange = residues 102-105); the residues involved in the interface are shown as red dots. The yellow line associated to some residues, describes the direction and the amplitude of the displacement described by PC1 during the trajectory for that residue.*

Also in this case the different starting conformations of the simulations play an important role. In 1CTQ_Au the motions are mainly localized in two regions, the residues 120-140 and the switch I. More interestingly, the analysis of the separated simulation in the switches region indicates as once the partner is removed the domain seems to "open" the interface pocket. In fact the direction of the motion in these regions (indicated by the yellow lines in Fig 4.12) indicates that the switch II is moving toward the p-loop and the switch I is changing in the opposite direction. On the contrary in the bound simulation, the PC1 mainly describes the motion of the residues 120-140 at the other side of the domain with respect to the binding interface.

The tendency of the switches to change the starting conformation of the complex once removed the partner, in the separated simulation, is evident also in the other cases where the switch region is involved in the interface (see Fig 4.4b-d). On the contrary, in these systems the bound form, both in the open and in the close state, shows a more constrained dynamics.

Once described the flexibility of the Ran, Rab and H-Ras domains, using both the RMSF and the PC1 analyses, the simulations were studied using the SOM approach developed in this thesis (Par 2.3). The simulations were described using the Cartesian coordinates of the Cα. The length of the sequences in these three cases is different (see the alignment in Fig. 4.3), producing a different length of the input data vectors presented to the SOM (Ran (1IBR_A) = 486, H-Ras = 489, Rab = 492 elements). As shown in Table 4.2 the simulations have different length, from 25 to 75 ns producing a different number of data vectors. Using the sampling rate defined in Par 2.3.5 of 1/100 ps the resulting number of conformation extracted were: 250 for the 25 ns simulations of, 500 for the 50 ns simulations, and 750 for the 75ns simulation.

First we analysed each single trajectory to describe the conformations that best summarize the conformational sampling of each simulation. Only the analyses of the single trajectories of H-Ras are reported, as an example, to present the general trends. The Mojena's rule (69) applied after clustering of the output SOM obtained in the analyses of 1CTQ_Au, 1BKD_Rs and 1BKD_Rb produced a division of each map in 4 clusters (Fig 4.13).

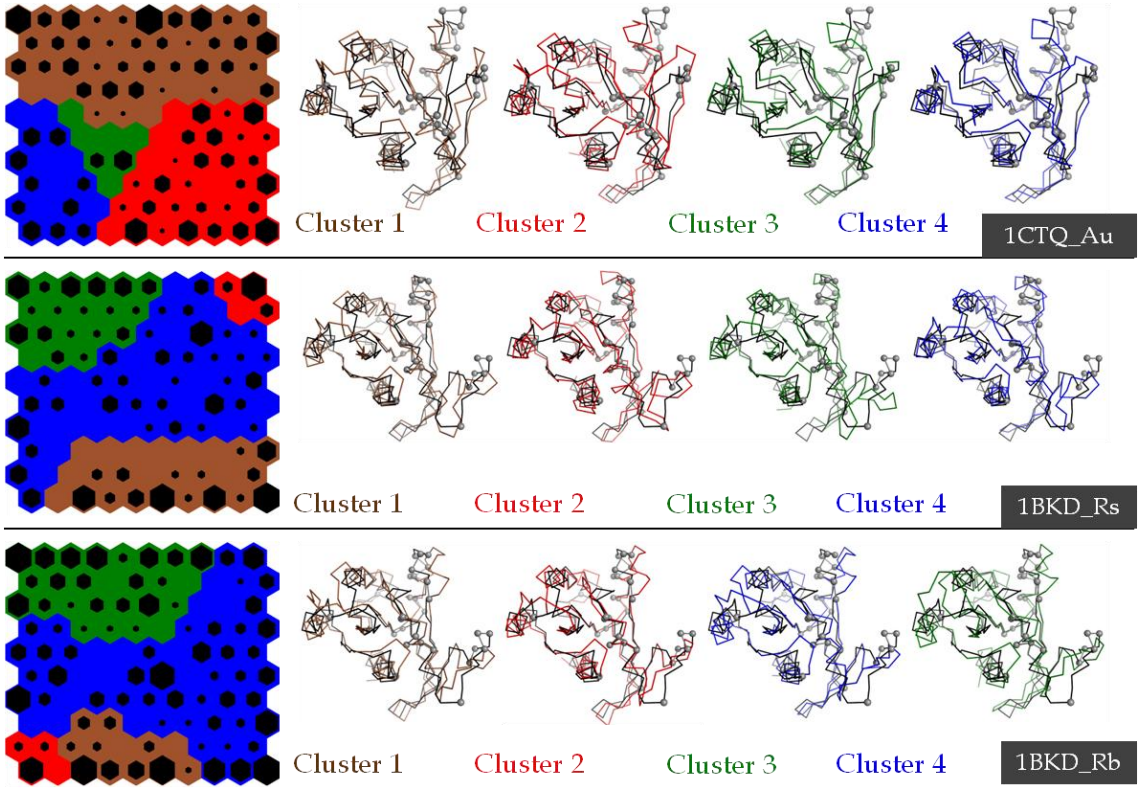



*Figure 4.13: H-Ras SOM analysis – In each section: on the left, the representation of the bidimensional output SOM (see Fig. 4.9 for the details. On the right, the ribbon representation of the conformation of the centroid of each cluster, superimposed with a reference structure (conformation at t=0 of the MD simulation), reported in black. The residues involved in the interface are highlighted as grey spheres.*

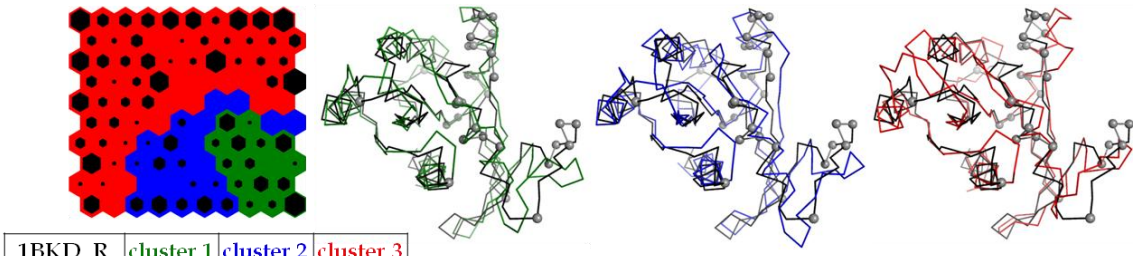
For the unbound simulation (1CTQ_Au) the previous analyses (RMSF and PC1) (Fig 4.11 and 4.12) suggested that the regions with the highest conformational changes are the switch region and residues 120-130. Analysing the centroids of the cluster obtained (upper part of Fig 4.13) this indication is confirmed: clusters 1 and 2 describe similar conformations with differences located in the switch I; clusters 3 and 4 have arrangements of the residues 120-130 different from 1 and 2 and differ for the arrangement of the switch I.

The SOM analysis of the separated simulation (1BKD_Rs) (middle part of Fig 4.13) is mainly driven by the big conformational changes of the switch I. Therefore all the four clusters are characterized by different arrangements of this region. But, with the exception of cluster 2, all the other clusters describe also a different

arrangement of the residue 130-140 with respect to the other two clusters. The conformations of these clusters allow us to clearly understand what suggested by the PC1 analysis (Fig 4.12): the conformations of the switch I describe a progressive opening of the switch I in the different clusters, with respect to the reference structure.

From the SOM analysis of the bound simulation (1BKD_Rb) (lower part of Fig 4.13) it emerges that the conformations of the cluster centroids describe a more constrained dynamics of the switch region, involved in binding. The regions with the most relevant conformational changes, that drive the learning process of the SOM, are the residues 102-105 and 120-130. Interestingly all the centroids, compared with the reference structure, do not show superimposition in the switch I region. Remembering that the reference structure is the structure of 1BKD_Rs at t=0 of the simulation (i.e. after the equilibrations steps of the MD protocol) this suggests that the conformational basins of the two simulations, bound and separated, have some differences.

This is confirmed by the SOM analysis of the combined separated and bound simulations (Fig 4.14). The analysis of the centroids allows to detect the specific conformations that best summarize the conformational sampling of each trajectory compared with the other. In fact each cluster obtained is populated by conformations of only one simulation: clusters 1 and 2 only by conformations of the separated simulations and cluster 3 only by conformations of the bound one. Clusters 1 and 2 re-confirm the analysis of the separated simulation, where the regions with the highest flexibility were the switch I and the residues 120-130. The centroid of cluster 3 confirms the different conformational arrangement of switch I and residues 120-130 during the bound simulation.



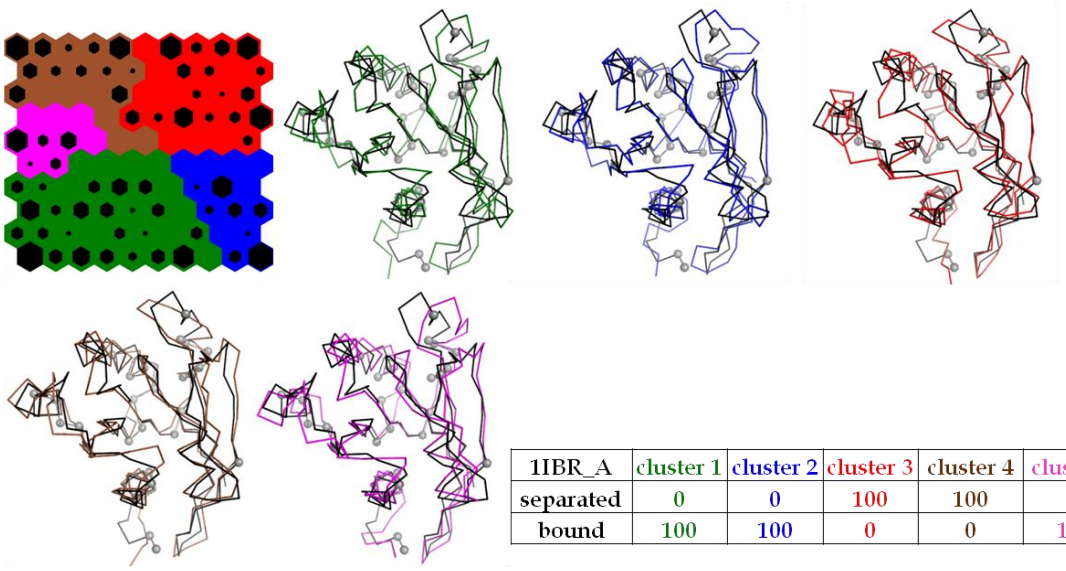

| 1BKD_R | cluster 1 | cluster 2 | cluster 3 |
|---|---|---|---|
| separated | 100 | 100 | 0 |
| bound | 0 | 0 | 100 |

*Figure 4.14: H-Ras SOM analysis: comparison between separated and bound simulations. On the left, the representation of the bidimensional output SOM (see Fig. 4.9 for the details) and, at bottom, the percentage distribution of conformations of each simulation in the clusters. On the right, the ribbon representation of the conformation of the centroid of each cluster, superimposed with a reference structure (conformation of 1BKD_Rs at t=0 of the MD simulation), reported in black. The residues involved in the interface are highlighted as grey spheres.*

Also in the other simulations the dynamical behaviour of the switches involved in the interfaces leads to a different conformational sampling between the bound and the separated simulations. This was confirmed by the analyses of the centroids of the clusters obtained with the SOM analyses of the single trajectories (data not shown) and it can be highlighted discussing the SOM results obtained with the combined analyses of the bounds and separated simulations.

The SOM obtained for Ran (1IBR_A) is divided in five clusters (Fig. 4.15). As a confirmation that different conformational basins were sampled during the two simulations, there are not shared clusters between the two simulations. Clusters 1, 2 and 5 describe the bound simulation, clusters 3 and 4 the separated one. An interesting feature of the map obtained is that the border between the clusters that describe the bound simulation and those that describe the separated one are composed by empty neurons. This means that there are no conformations described by that space of information. Clusters 1, 2 and 5 mainly differ from clusters 3 and 4 by the conformation of the switch II. Within these two groups of clusters there are differences that reconfirms the characteristics of each simulation (not shown): clusters 1, 2 and 5 contain different arrangement of the switch II and cluster 5 differ to clusters 1 and 2 for a different conformations of the residues 130-

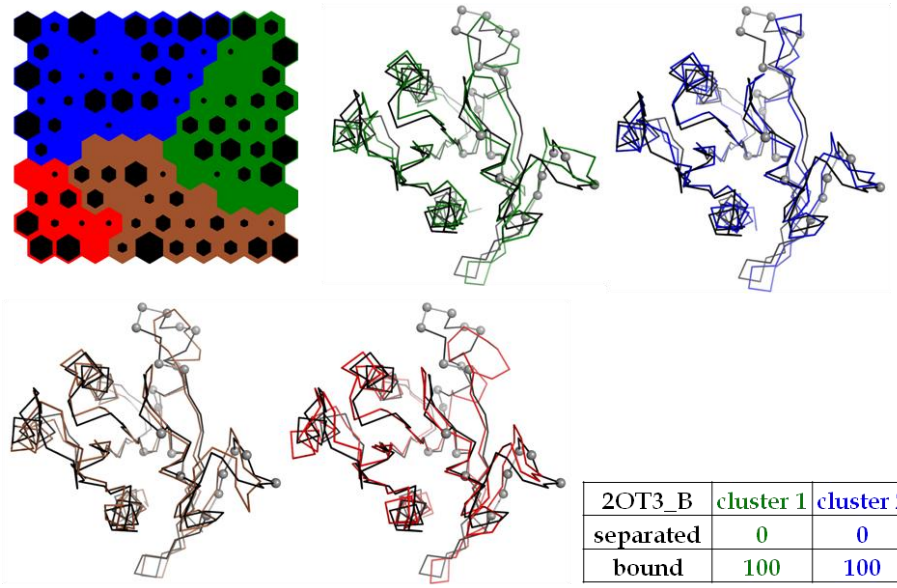
140, differences that were observed in the bound simulation; clusters 3 and 4 describe the different conformations of the residues 102-105 and 120-140, sampled during the separated simulations.



| 1IBR_A | cluster 1 | cluster 2 | cluster 3 | cluster 4 | cluster 5 |
|---|---|---|---|---|---|
| separated | 0 | 0 | 100 | 100 | 0 |
| bound | 100 | 100 | 0 | 0 | 100 |

*Figure 4.15: Ran (1IBR_A) SOM analysis: comparison between separated and bound simulations. On the left, the representation of the bidimensional output SOM (see Fig. 4.9 for the details. On the right, the ribbon representation of the conformation of the centroid of each cluster, superimposed with a reference structure (conformation of 1IBR_As at t=0 of the MD simulation), reported in black. The residues involved in the interface are highlighted as grey spheres. At bottom, the percentage distribution of conformations of each simulation in the clusters.*

The SOM analysis of Rab (Fig. 4.16) produced 4 clusters and, once again, none of them shared conformations within both of the simulations. In fact clusters 1 and 2 contain conformations of the bound simulation and clusters 3 and 4 of the separated one. Moreover, also in this case there is the presence of empty neurons in the border between the clusters of the bound and of the unbound simulations.
Also in this complex the switch II is involved in the interface (Fig 4.4d) and as indicated by the RMSF plot (Fig 4.11) this region is the one with the highest flexibility. This evidence is confirmed by the SOM analysis, as the conformations of each centroid permit to identify different conformations of the switch II during the two simulations.



| 2OT3_B | cluster 1 | cluster 2 | cluster 3 | cluster 4 |
|---|---|---|---|---|
| separated | 0 | 0 | 100 | 100 |
| bound | 100 | 100 | 0 | 0 |

*Figure 4.16: Rab SOM analysis: comparison between separated and bound simulations. On the left, the representation of the bidimensional output SOM (see Fig. 4.9 for the details. On the right, the ribbon representation of the conformation of the centroid of each cluster, superimposed with a reference structure (conformation of 2OT3_Bs at t=0 of the MD simulation), reported in black. The residues involved in the interface are highlighted as grey spheres. At bottom, the percentage distribution of conformations of each simulation in the clusters.*

### *4.4 MD simulations and analysis of the effectors dynamics*

The simulations used for this analysis are 1I2M_Bs and 1I2M_Bb for RCC1, 1IBR_Bb and 1IBR_Bs for Importin β, 1BKD_Ss and 1BKD_Sb for SOS-1 and 2OT3_As and 2OT3_Ab for Rebex-5 (see Table 4.1).

The study of the flexibility of these domains were not in the main aim of this study, but by a methodological point of view it was interesting to verify the exportability of the SOM protocol developed here to domains with different dimension. The dimensions of these domains are: 394 residues for RCC1, 456 residues for Importin β, 477 residues for SOS-1 and 253 residues for Rebex-5.

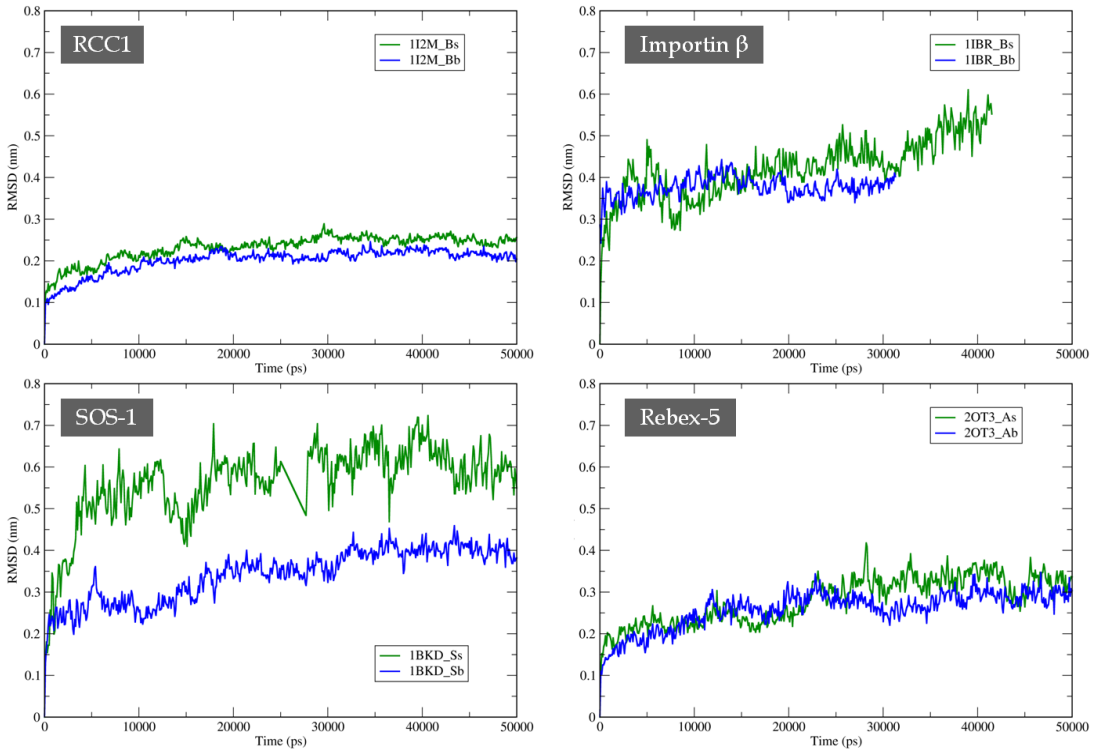



*Figure 4.17: RMSD computed on Ca. For each case, in green the "separated" simulation and in blue the "bound" simulation. For each trajectory, the structure used as reference to compute the RMSD values is the conformation of the system at t=0 of the simulation.*

Concerning the results of the MD simulations, the plots of the RMSD values from the starting structure of the simulation, computed on the Cα, versus the simulation time are reported in Fig. 4.17. The average deviation observed in these cases changes from 0.2 nm in the RCC1 simulation to 0.6 nm in that of SOS-1. Also the comparison between the bound and the separated simulations is not univocal. In two cases (Importin β and Rebex-5) the average deviation is the same in the two simulations. In RCC1 a slight increase is observed in the separated and this difference is more evident in the SOS-1 simulation in which the separated simulation is characterized by values between 0.4 and 0.7 nm, while the bound one by values in the range 0.2 – 0.4 nm. In all the graphs all the simulations are drifting toward higher values of RMSD suggesting a possible poor convergence of the sampling.

To verify the completeness of conformational sampling, the overlap between the conformational space spanned in each half of the simulation and that of the overall simulation was calculated, as described in Par 2.1.4. The result in Table 4.4 confirm the indications given by the RMSD plot. In fact in all the cases the overlap between one half and the whole trajectory is below 0.6. Moreover, in general the overlap of the first half is higher (0.59 – 0.71) than the overlap in the second half (0.46 – 0.68).

*Table 4.4: Overlap of sampling in the MD simulations of the effectors domains: overlap between the conformational space spanned by each half of the effectors simulation and that of the overall trajectory. In the last column the length of the simulation.*

| Domain | Case | 1st half | 2nd half | Time (ns) |
|---|---|---|---|---|
| RCC1 | 1I2M_Bs | 0.68 | 0.53 | 50 |
| | 1I2M_Bb | 0.71 | 0.52 | 50 |
| Importin β | 1IBR_Bs | 0.65 | 0.54 | 50 |
| | 1IBR_Bb | 0.59 | 0.68 | 50 |
| SOS-1 | 1BKD_Ss | 0.71 | 0.53 | 50 |
| | 1BKD_Sb | 0.63 | 0.46 | 50 |
| Rebex-5 | 2OT3_As | 0.59 | 0.53 | 50 |
| | 2OT3_Ab | 0.66 | 0.56 | 50 |

To extract the most informative directions of motion the Essential Dynamics analysis (see Par. 2.1.5) was performed. The fraction of total motion described by different subspaces was evaluated to identify the extent of the essential space for each domain. The results reported in Table 4.5 indicate that, in all the cases, more than 70% of the motion is described by the first 15 eigenvectors. Considering the dimensions of the systems, this means that it is possible to describe the 70% of the information of the simulations by using less than 1% of the total space. Therefore a 15 dimensional essential space was selected for the following analyses.

*Table 4.5: Effectors Essential Space: Percentage of total space described using increasing number of eigenvectors*

| | RCC1 | | Importin β | | SOS-1 | | Rebex-5 | |
|---|---|---|---|---|---|---|---|---|
| PC | 1I2M_Bb | 1I2M_Bs | 1IBR_Bb | 1IBR_Bs | 1BKD_Sb | 1BKD_Ss | 2OT3_Ab | 2OT3_As |
| 1 | 29.8 | 27.7 | 36.1 | 38.3 | 46.2 | 38.5 | 30.8 | 47.2 |
| 1-2 | 39.5 | 37.4 | 55.8 | 56.3 | 58.4 | 51.9 | 46.7 | 58.9 |
| 1-3 | 45.5 | 43.0 | 62.4 | 66.0 | 66.9 | 62.5 | 54.1 | 64.3 |
| 1-5 | 52.1 | 52.1 | 71.6 | 75.1 | 73.6 | 70.9 | 62.8 | 71.3 |
| 1-10 | 61.2 | 62.6 | 80.7 | 83.7 | 81.3 | 80.7 | 73.3 | 79.4 |
| **1-15** | **70.1** | **71.5** | **84.5** | **87.8** | **84.8** | **85.0** | **78.5** | **83.6** |
| 1-20 | 80.6 | 82.1 | 86.8 | 89.9 | 87.0 | 87.4 | 81.9 | 86.2 |
| 1-50 | 87.9 | 92.9 | 92.3 | 94.5 | 92.8 | 92.9 | 90.5 | 92.9 |
| 1-all | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |

Using the same structural representation used until now, i.e. the Cartesian coordinates of the Cα, the number of elements in the input data vectors for the SOM is significantly higher compared with the case used ad test (SH3 domain, 55 residues corresponding to 165-dimensional vectors). The corresponding length of the vectors is 1182 elements for RCC1, 1368 elements for Importin β, 1431 elements for SOS-1 and 759 elements for Rebex-5.

For each domain, both the SOM analyses of the single trajectories and the analyses of the combined separated and bound simulations were performed. In the analysis of the single trajectories we were able to define the conformations that best summarize each trajectory. As examples, in Fig 4.18 the SOM results for 1BKD_Ss, and in Fig 4.19 the results for the combined simulations for 1BKD_S are reported.

In general, the SOM analyses indicate that the region involved in the interface shows the highest difference between the bound and the separated dynamics. Moreover, also other regions not directly involved in the interface show relevant differences when their motion are correlated with that at the interface.

As shown for the example in Fig. 4.19 (1BKD_S), the comparison between the separated and the bound simulations in all the studied cases indicates that the conformational spaces sampled in the bound and in the separated trajectories do not show common conformations.

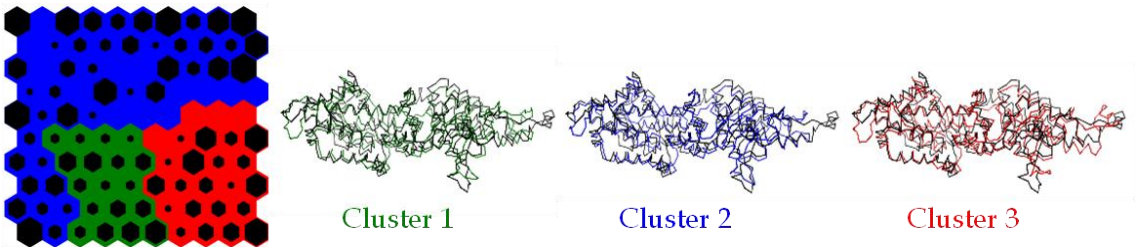
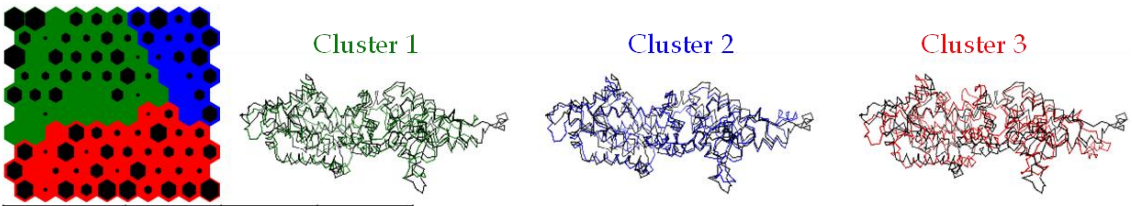



*Figure 4.18: SOS-1 SOM analysis: On the left, the representation of the bidimensional output SOM (see Fig. 4.9 for the details). On the right, the ribbon representation of the conformation of the centroid of each cluster, superimposed with a reference structure (conformation of 1BKD_Ss at t=0 of the MD simulation), reported in black.*




| 1BKD_S | cluster 1 | cluster 2 | cluster 3 |
|---|---|---|---|
| separated | 100 | 100 | 0 |
| bound | 0 | 0 | 100 |

*Figure 4.19: SOS-1SOM analysis: comparison between separated and bound simulations. On the left, the representation of the bidimensional output SOM (see Fig. 4.9 for the details) and, at bottom, the percentage distribution of conformations of each simulation in the clusters. On the right, the ribbon representation of the conformation of the centroid of each cluster, superimposed with a reference structure (conformation of 1BKD_Ss at t=0 of the MD simulation), reported in black.*

### *4.5 Conclusions*

As described in Par 4.2.2, the MD simulations performed for the Ras proteins were strongly influenced by the large structural differences between the starting crystal structures of the bound and the unbound systems. In fact, it resulted that these simulations were not able to sample the conformational changes between the starting structures. This first evidence was also confirmed by the SOM analysis where, due to the differences in the two conformational basins spanned, the bound and unbound trajectories could not be compared by a unique combined map. As a consequence, the unbound simulations were not used for further comparisons.

The MD results of the bound and separated simulations, on the contrary, indicated that a direct comparison between the two conformational ensembles was possible. These simulations started from the same conformation and the artificial separation from the effectors had the aim to study the differences, in terms of flexibility, due to its presence or absence.

In both the RMSD and the RMSF analyses (Par. 4.2.2 – 4.3.2) no general relative behaviours between these simulations emerged. From these results it was expected that the SOM analysis could highlight both the conformational space common to the bound and separated systems and the specific space sampled by the Ras protein only in the presence of the effector. This was confirmed only in one case, Ran (1I2M_A) (Par. 4.3.1). Analyzing the specific dynamical behaviour of this domain, this seems to be connected with the fact that the binding interface only partially involves the switch region. In this case the comparison between the bound and the separated simulations produced a SOM with a cluster containing conformation from both the trajectories. Interestingly, the shared cluster contains conformations sampled during the first nanoseconds of the simulations; after

that, the presence/absence of the partner led the simulations toward different conformational basins.

In all the other cases (Par 4.3.2) the arrangement of the switches seems to be strongly regulated by the interactions with the partner. In these cases, in fact, the interface mostly involves the switch region. Once the partner is removed, the conformations of the switches undergo large conformational changes, that make the trajectories not directly comparable. The equilibration stage of the simulations (Par 4.2.1) was enough to allow the domain to re-arrange this region.

The features of the SOMs obtained in these cases highlighted this behaviour (Fig 4.14 - 4.16). First, the clustering of the maps did not produce any shared clusters between the separated and bound simulations; the conformational analysis of the centroids of the clusters allowed us to detect the differences that made the comparison not possible. Second, the maps were clearly divided in two regions, one containing the conformations of the separated simulation and one the conformations of the bound one. Within these two regions the clusters described the most interesting conformations of each simulation but the borders between the two regions were mainly composed by empty neurons (see maps in Fig 4.14 – 4.16). This fact must be related with the neighbourhood property of the SOM learning process. The presence of empty neurons between two regions means that the information learned by each regions is so different from the other that the space described by the neurons in the middle does not describe any real data.

Both the MD results and the SOM analyses confirmed that these preliminary MD simulations were not able to give the conformational sampling needed to study the role of flexibility in the formation of these complexes.

Finally, concerning the exportability of the SOM protocol developed with the SH3 domain, these analyses gave some useful insight.

Using the same parameters set for a case with a dimension of 55 residues, the SOM approach was appropriate to analyze both the Ras proteins (c.a. 160 residue) and the effectors (from 250 to 480 residues). In particular, this means that a map composed by 100 neurons is sufficiently wide to describe systems up to ten times bigger compared with the test case.

It can be concluded that the SOM approach could be used to compare the trajectories of the Ras proteins in the bound and unbound forms. However SOM is just an analysis of what has been sampled and cannot predict essential motions that would lead to conformational changes. Therefore appropriate strategies have to be studied to more effectively sample the conformational landscape, overcoming the energy barriers among minima.

# Chapter 5

## CONCLUSIONS

We studied a novel approach to compare conformational ensembles of protein domains with the goal of highlighting similarities and differences in functional motions.

The novelty of the approach concerns the development of a methodology to analyse data obtained via MD simulations with a specific neural network, the Self-Organizing Maps (SOMs), that has recently been shown to be suitable for the analysis of individual MD trajectories (22).

When dealing with large datasets of conformations, a major issue is the computational cost of the analysis. Two possible strategies to overcome this problem are to compare only the average geometrical properties of subgroups of data or to apply a two-stage selection.

An example of the former strategy is the comparison of the ensemble RMSF on atom positions of functionally related proteins. While the analysis is relatively informative and fast, small fluctuations are difficult to detect, the direction of motion associated to each peak is not considered, and comparisons are only pair-wise. These limitations clearly appeared in our analysis of the RMSF profiles of the study-cases (Par 3.2.2 and 4.3). On the contrary, the results presented in this thesis highlighted that the proposed SOM approach retains high sensitivity, is able to differentiate motions with similar average fluctuation and is not restricted to pairwise comparisons.

The second strategy involves using geometrical clustering methods in a two-stage or sieved approach (22) by initially clustering only part of the data and then in a second step by adding the missing ones into existing clusters. This decreases

the computational time significantly but can lead to the loss or distortion of the topological relations among the original data and eventually to a biased grouping, if the selection at first stage is not representative. Similarly during a SOM training each data vector is compared only to the neuron vectors representing all the data already presented to the map, but the topological relations are intrinsically recorded and the representative geometries are dynamically updated avoiding a bias. These advantages come at increased computational cost in the training stage. However, once a map is trained on a group of representative protein domains, it can be used for fast classification of conformational ensembles of similar systems.

As described in Chapters 3 and 4, the protocol here developed was applied to two study cases. In the first case we were interested in studying the modulation of the flexibility by mutagenesis, in the second one the role of flexibility in protein binding.

To study the effects of mutagenesis on a domain flexibility we selected the Src-SH3 domain and its six mutants as a test case (Chapter 3). The same systems had also been used to develop and test the method (59). To test the performance of the SOM approach we studied the systems at three levels: single trajectories, pairs of trajectories, the whole Src-SH3 data set. At all the levels the method showed its ability in the extraction of the flexibility information (Par. 3.3). In the analyses of the single trajectories it effectively described the most relevant conformations sampled during each MD simulations. Interestingly the representative conformations extracted from each map after clustering of the neurons were able to describe both cases in which large conformational changes occurred during the simulation (e.g. the R21G mutant) and cases in which the simulation had opposite behavior (e.g the wild-type SH3 domain). By analysing multiple simulations the SOM approach was able to detect similarities and differences in flexibility of the mutants with high sensitivity. Moreover, the analysis of the representative

conformations in the clustered map confirmed a relation between the dynamical behavior of the domain, associated with the specific point-mutation, and the biological function, i.e. the binding affinity with the p41 decapeptide.

To study the role of flexibility in protein binding we selected a group of complexes in which one of the partners belongs to the Ras superfamily of sociable hubs proteins (Chapter 4). The dynamics of the domains was studied via MD simulations in three different conditions: in presence/absence of the partner (bound and separated forms), and starting from a different crystal structure (unbound forms). The MD simulations performed were insufficient to sample the large conformational changes occurring between the starting structures of the bound and the unbound Ras systems, that mainly involves the "switch region" (Par. 4.1). Therefore the two trajectories could not be compared by a unique combined SOM. On the contrary, it was possible to study the effect of the presence/absence of the partner by comparing the bound and separated simulations in one of the system studied (the Ran domain in complex with the RCC1), in which the binding interface only partially involves the switch region. Using the SOM approach it was possible to detect both the common conformational space sampled during the two simulations and the peculiar conformations of each simulation, thus detecting the dynamical features connected to the binding to the effector (Par. 4.3.1). In the other complexes, in which the interface mostly involves the switch region, this comparison was not possible. In fact the simulations with and without the partner diverged, starting from the first steps of the MD simulations, in two different basins, and this caused an output map divided in two regions with no overlap between them. In these cases we have to consider that the length of the simulations could be not enough to obtain a convergent conformational sampling, or that different techniques could be more appropriate to effectively sample these conformational landscapes characterized by high energy barriers among minima. These hypotheses are under study.

By a methodological point of view, the analysis of the Ras complexes was also useful to test the exportability of the protocol developed with the Src-SH3 domain to domains of different size. In particular, the SOM protocol was successfully applied the trajectories of the effectors of Ras proteins (Par. 4.4), demonstrating that a SOM composed by 100 neuron is able to describe the conformational changes of domains up to c.a. ten times the Src-SH3 domain.

The analysis of the results obtained in this thesis underlines additional interesting aspects regarding the application of the proposed SOM approach that deserve specific attention and suggest future research directions.

An interesting feature of the output SOMs that clearly emerged from the SH3 study case (Chapter 3) is that the conformational transitions during the MD simulations are indirectly recorded on the map: neurons describing conformations involved in a transition are adjacent on the map. This is not a trivial outcome because the simulation time of a given conformation is not given as input information to the SOM. This feature is connected with the preservation of the topology of the data in the final map, obtained by using the neighbourhood properties during the training phase. If the extension of the protocol to several study cases will confirm the generality of this observation, a future direction will be the use of a trained map for fast stochastic simulations.

The approach here presented is independent from the method used to generate the structural ensemble and is reliable to describe both small and large differences. It is therefore suitable to also analyse combinations of ensembles from computational methods with a more extended sampling of the conformational space and from experiments (NMR ensembles or multiple X-ray depositions of the same structure).
The test-cases here presented regard the comparison of a same domain in different forms (with or without an associated partner protein) or of mutants of a

single domain, that can easily be aligned. A future development of this study is the identification of alternative representations of protein conformations, that do not require the preliminary definition of structurally equivalent positions by structural alignment. This will allow an extension on the comparison of different domains, including distant homologous proteins.

# BIBLIOGRAPHY

1. **Henzler-Wildman K., Kern D.,** *Dynamic personalities of proteins.*, Nature , 2007, Vol. 450, pp. 964-972.

2. **Boehr D.D., Nussinov R., Wright P.E.,** *The role of dynamic conformational ensembles in biomolecular recognition.,* Nat Chem Biol, 2009, Vol. 5, pp. 789-796.

3. **Berendsen H.J., Hayward S.,** *Collective protein dynamics in relation to function.,* Curr Opin Struct Biol, 2000, Vol. 10, pp. 165-169.

4. **Karplus M., Kuriyan J.,** *Molecular dynamics and protein function,* Proc Natl Acad Sci U S A, 2005, Vol. 102, pp. 6679-6685.

5. **Hub J.S., de Groot B.L.,** *Detection of functional modes in protein dynamics,* PLoS Comput Biol, 2009, Vol. 5, p. e1000480.

6. **Huber R., Bennett Jr W.S.,** *Functional significance of flexibility in proteins,* Biopolymers, 1983, Vol. 22, pp. 261–279.

7. **Dobbins S.E., Lesk V.I., Sternberg M.J.E.,** *Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking,* Proc Natl Acad Sci U S A, 2008, Vol. 105, pp. 10390-10395.

8. **Higurashi M., Ishida T., Kinoshita K.,** *Identification of transient hub proteins and the possible structural basis for their multiple interactions,* Protein Sci, 2008, Vol. 17, pp. 72–78.

9. **Tokuriki N., Tawfik D.S.,** *Protein dynamism and evolvability,* Science, 2009, Vol. 324, pp. 203-207.

10. **Maguid S., Fernandez-Alberti S., Echave J.,** *Evolutionary conservation of protein vibrational dynamics.,* Gene, 2008, Vol. 422, pp. 7-13.

11. **Pandini A., Mauri G., Bordogna A., Bonati L.,** *Detecting similarities among distant homologous proteins by comparison of domain flexibilities,* Protein Eng Des Sel, 2007, Vol. 20, pp. 285-299.

12. **Maguid S., Fernandez-Albert S.i, Parisi G., Echave J.,** *Evolutionary conservation of protein backbone flexibility,* J Mol Evol, 2006, Vol. 63 , pp. 448-457.

13. **Pandini A., Bonati L.,** *Conservation and specialization in PAS domain dynamics,* Protein Eng Des Sel, 2005, Vol. 18, pp. 127-137.

14. **Cascella M., Micheletti C., Rothlisberger U., Carloni P.,** *Evolutionarily conserved functional mechanics across pepsin-like and retroviral aspartic proteases,* J Am Chem Soc, 2005, Vol. 127, pp. 3734-3742.

15. **Velázquez-Muriel J.A., Rueda M., Cuesta I., Pascual-Montano A., Orozco M., Carazo J.M.,** *Comparison of molecular dynamics and superfamily spaces of protein domain deformation,* BMC Struct Biol, 2009, Vol. 9, p. 6.

16. **Pang A., Arinaminpathy Y., Sansom M.S., Biggin P.C.,** *Comparative molecular dynamics--similar folds and similar motions?,* Proteins, 2005, Vol. 61, pp. 809-822.

17. **Scheraga H.A., Khalili M., Liwo A.,** *Protein-folding dynamics: overview of molecular simulation techniques,* Annu Rev Phys Chem, 2007, Vol. 58, pp. 57–83.

18. **van Gunsteren W.F., Bakowies D., Baron R., Chandrasekhar I., Christen M., Daura X., Gee P., Geerke D.P., Glättli A., Hünenberger P.H., Kastenholz M.A., Oostenbrink C., Schenk M.,**

**Trzesniak D., van der Vegt N.F., Yu H.B.,** *Biomolecular modeling: Goals, problems, perspectives,* Angew Chem Int Ed Engl, 2006, Vol. 45, pp. 4064-4092.

19. **Hess B.,** *Convergence of sampling in protein simulations,* Phys Rev E, 2002, Vol. 65, p. 031910.

20. **van der Kamp M.W., Schaeffer R.D., Jonsson A.L., Scouras A.D., Simms A.M., Toofanny R.D., Benson N.C., Anderson P.C., Merkley E.D., Rysavy S., Bromley D., Beck D.A., Daggett V.,** *Dynameomics: a comprehensive database of protein dynamics,* Structure, 2010, Vol. 18, pp. 423-435.

21. **Meyer T., D'Abramo M., Hospital A., Rueda M., Ferrer-Costa C., Pérez A., Carrillo O., Camps J., Fenollosa C., Repchevsky D., Gelpí J.L., Orozco M.,** *MoDEL (Molecular Dynamics Extended Library): A Database of Atomistic Molecular Dynamics Trajectories,* Structure, 2010, Vol. 18, pp. 1399-1409.

22. **Shao J., Tanner S.W., Thompson N., Cheatham T.E.,** *Clustering molecular dynamics trajectories: Characterizing the Performance of different clustering algorithms,* J Chem Theory Comput, 2007, Vol. 3, pp. 2312-2334.

23. **Chodera J.D., Singhal N., Pande V.S., Dill K., Swope W.,** *Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics,* J Chem Phys, 2007, Vol. 126, p. 155101.

24. **Noé F., Horenko I., Schütte C., Smith J.C.,** *Hierarchical Analysis of Conformational Dynamics in Biomolecules: Transition Networks of Metastable States,* J Chem Phys, 2007, Vol. 126, p. 155102.

25. **Muff S., Caflisch A.,** *Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a b-sheet miniprotein,* Proteins, 2008, Vol. 70, pp. 1185-1195.

26. **Huisinga W., Best C., Roitzsch R., Schütte C., Cordes F.,** *From Simulation Data to Conformational Ensembles: Structure and Dynamics based Methods,* J Comput Chem, 1999, Vol. 20, pp. 1760-1774.

27. **Deuflhard P., Huisinga W., Fischer A., Schutte C.,** *Identification of Almost Invariant Aggregates in Reversible Nearly Uncoupled Markov Chains,* Linear Alg Appl, 2000, Vol. 315, pp. 39-59.

28. **Keller B., Daura X., van Gunsteren W.F.,** *Comparing geometric and kinetic cluster algorithms for molecular simulation data,* J Chem Phys, 2010, Vol. 132, p. 074110.

29. **Karpen M.E., Tobias D.J., Brooks C.L.,** *Statistical clustering techniques for the analysis of long molecular dynamics trajectories: analysis of 2.2-ns trajectories of YPGDV,* Biochemistry, 1993, Vol. 32, pp. 412-420.

30. **Shenkin P.S., McDonald D.Q.,** *Cluster analysis of molecular conformations,* J Comput Chem, 1994, Vol. 15, pp. 899-916.

31. **Torda A.E., van Gunsteren W.F.,** *Algorithms for clustering molecular dynamics configurations,* J Comput Chem, 1994, Vol. 15, pp. 1331-1340.

32. **DeFanti T.A., Brown M.D.,** *Visualization expanding scientific and engineering research opportunities, Readings in Information Visualization Using Vision to Think;* Card S.K., Mackinlay J.D., Schneiderman B., 1989, pp. 37–56.

33. **Kohonen T.,** *The self-organizing map,* Proceedings of the Institute of Electrical and Electronics Engineers, 1990, Vol. 78, pp. 1464-1480.

34. **Hyvönen M.T., Hiltunen Y., El-Deredy W., Ojala T., Vaara J., Kovanen P.T., Ala-Korpela M.,** *Application of self-organizing maps in conformational analysis of lipids,* J Am Chem Soc, 2001, Vol. 123, pp. 810-806.

35. **Murtola T., Kupiainen M., Falck E., Vattulainen I.,** *Conformational analysis of lipid molecules by self-organizing maps,* J Chem Phys, 2007, Vol. 126, p. 054707.

36. **Bouvier G., Evrard-Todeschi N., Girault J.P., Bertho G.,** *Automatic clustering of docking poses in virtual screening process using self-organizing map,* Bioinformatics, 2010, Vol. 26, pp. 53-60.

37. **Douglas C., Montgomery D.C.** *Design and Analysis of Experiments, Student Solutions Manual.* Wiley, 2005.

38. **Casares S., López-Mayorga O., Vega M.C., Cámara-Artigas A., Conejero-Lara F.,** *Cooperative propagation of local stability changes from low-stability and high-stability regions in a SH3 domain,* Proteins, 2007, Vol. 67, pp. 531-547.

39. **Berendsen H.J.C., van der Spoel D., van Drunen R.,** *GROMACS: A message-passing parallel molecular dynamics implementation,* Comp Phys Comm, 1995, Vol. 91, pp. 43–56.

40. **Darden T., York D., Pedersen L.,** *Particle mesh Ewald: An N·log(N) method for Ewald sums in large Systems,* J Chem Phys, 1993, Vol. 98, pp. 10089–10092.

41. **van der Spoel D., Lindahl E., Hess B., van Buuren A.R., Apol E., Meulenhoff P.J.,Tieleman D.P., Sijbers A.L.T.M, Feenstra K.A., van Drunen R.,Berendsen H.J.C.,** *Gromacs User Manual version 4.0, www.gromacs.org,* 2005.

42. **Lindahl E., Hess B., van der Spoel D.,** *Gromacs 3.0: A package for molecular simulation and trajectory analysis,* J Mol Mod, 2001, Vol. 7, pp. 306–317.

43. **van der Spoel D., Lindahl E., Hess B., Groenhof G., Mark A.E., Berendsen H.J.C.,** *GROMACS: Fast, Flexible and Free,* J Comp Chem, 2005, Vol. 26, pp. 1701–1718.

44. **Hess B., Kutzner C., van der Spoel D., Lindahl E.,** *GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation,* J Chem Theory Comput, 2008, Vol. 4, pp. 435-447.

45. **van Gunsteren W.F., Billeter S.R., Eising A.A., Hünenberger P.H., Krüger P., Mark A.E., Scott W.R.P., Tironi I.G.,** *Biomolecular Simulation: The GROMOS96 Manual and User Guide,* Zürich : vdf Hochschulverlag AG an der ETH Zürich and BIOMOS, Groningen, 1996.

46. **W.R.P. Scott, P.H. Huenenberger, I.G. Tironi, A.E. Mark, S.R. Billeter, J. Fennen, A.E. Torda, T. Huber, P. Krueger, W.F. van Gunsteren.,** *The GROMOS Biomolecular Simulation Program Package,* J Phys Chem A, 1999, Vol. 103, pp. 3596-3607.

47. **Schulera L.D., van Gunsteren W.F.,** *On the Choice of Dihedral Angle Potential Energy Functions for n-Alkanes,* Mol Sim, 2000, Vol. 25, pp. 301-319.

48. **Daura X., Mark A.E., van Gunsteren W.F.,** *Parametrization of aliphatic CHn united atoms of GROMOS96 force field,* J Comp Chem, 1998, Vol. 19, pp. 535–547.

49. **Berendsen H.J.C., Postma J.P.M., van Gunsteren W.F., Hermans J.**, *Intermolecular Forces,* Pullman B. Dordrecht, Reidel 1981, pp. 331–342.

50. **Miyamoto S., Kollman P.A.,** *Settle: An analytical version of the shake and Rattle algorithms for rigid water models,* J Comp Chem, 1992, Vol. 13, pp. 952-962.

51. **Ryckaert J.P., Ciccotti G., Berendsen H.J.C.,** *Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes,* J Comput Phys, 1977, Vol. 23, pp. 327–341.

52. **Hess B., Bekker H., Berendsen H.J.C., Fraaije J.G.E.M.,** *LINCS: A linear constraint solver for molecular simulations,* J Comp Chem, 1997, Vol. 18, pp. 1463–1472.

53. **Hockney R.W.,** *The potential calculation and some applications,* Meth Computl Phys, New York (1970), Vol. 9, pp. 135-211.

54. **Cuendet M.A., van Gunsteren W.F.,** *On the calculation of velocity-dependent properties in molecular dynamics simulations using the leapfrog integration algorithm,* J Chem Phys, 2007, Vol. 127, p. 184102.

55. **Amadei A., Linssen A.B.M., Berendsen H.J.C.,** *Essential dynamics of proteins,* Proteins, 1993, Vol. 17, pp. 412–425.

56. **Hertz J., Krogh A., Palmer R.G.** *Introduction to the theory of neural computation.* Addison-Wesley, 1991.

57. **Hagan M.T., Demuth H.B., Beale M.** *Nerual network design.* PWS Pub. Co., 1996.

58. **Vesanto J.,** *SOM-based data visualization methods,* Intell Data Anal, 1999, Vol. 3, pp. 111-126.

59. **Fraccalvieri D., Pandini A., Stella F., Bonati L.,** *Conformational and functional analysis of molecular dynamics trajectories by Self-Organising Maps,* under review, BMC Bioinformatics.

60.**Chan C.K.K., Hsu A.L., Tang S.L., Halgamuge S.K.,** *Using Growing Self-Organizing Maps to Improve the Binning Process in Environmental Whole-Genome,* J Biomed Biotechnol, 2008, p. 513701.

61. **Newman A.M., Cooper J.B.,** *AutoSOME: a clustering method for identifying gene expression modules without prior knowledge of cluster number,* BMC Bioinformatics, 2010, Vol. 11, p. 117.

62. **Clarke B., Fokoué E., Zhang H.H.,** *Principles and theory for data mining and machine learning,* Springer, Dordrecht, 2009.

63. **Roy R.K.,.** Design of Experiments Using the Taguchi Approach. Wiley-IEEE, 2001, pp. 207–235 and 369–402.

64. **Raskutti B., Leckie C.,** *An Evaluation of Criteria for Measuring the Quality of Clusters,* Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, 1999, pp. 905-910.

65. **Myers R.H., Montgomery D.C.,** *Response Surface Methodology: Process and Product Optimization Using Designed Experiments,* Wiley Series in Probability and Statistics, 1995,.

66. *JMP: Version 7.* **Inc., SAS Institute.** NC, 1989-2007, Cary.

67. **Cressie N., Read T.R.C.,** *Multinomial goodness-of-fit tests,* J R Stat Soc Ser B, 1984, Vol. 46, pp. 440-464.

68. **Xu R., Wunsch II D.C.,** *Clustering,* Hoboken. New Jersey : John Wiley and Sons, 2009.

69. **R., Mojena.,** *Hierarchical grouping methods and stopping rules: An evaluation,* Comput J, 1977, Vol. 4, pp. 359-363.

70. *Statistics Toolbox 7 User's Guide.* **M.A., Natick.** 1993-2009, The Mathworks Inc.

71. **Musacchio A., Noble M., Pauptit R., Wierenga R., Saraste M.,** *Crystal structure of a Src-homology 3 (SH3) domain,* Nature, 1992, Vol. 359, pp. 851-855.

72. **Fernandez-Ballester G., Beltrao P., Gonzales J.M., Song Y., Wilmanns M., Valencia A., Serrano L.,** *Structure-based prediction of the saccharomyces cerevisiae SH3-ligand interactions,* J Mol Biol, 2009, Vol. 388, pp. 902-916.

73. **Li S.S.,** *Specificity and verstaility of SH3 and other proline-recognition domanis: structural basis and implications for cellular signal transduction,* Biochem J, 2005, Vol. 390, pp. 641-653.
74. **Wang C., Pawley N.H., Nicholson L.K.,** *The role of backbone motions in ligand binding to the c-Src SH3 domain,* J Mol Biol, 2001, Vol. 313, pp. 873-887.
75. **Casares S., Ab E., Eshuis H., López-Mayorga O., van Nuland N.A.J., Conejero-Lara F.,** *The high-resolutionNMR structure of the R21A Spc-SH3:P41 complex: Understanding the determinants of binding affinity by comparison with Abl-SH3,* BMC Struct Biol, 2007, p. 7:22.
76. **Dominguez C., Boelens R., Bonvin A.M.,** *HADDOCK: a protein-protein docking approach based on biochemical or biophysical information,* J Am Chem Soc, 2003, Vol. 125, pp. 1731-1737.
77. **Vega M.C., Martinez J.C., Serrano L.,** *Thermodynamic and structural characterization of Asn and Ala residues in the disallowed II' region of the Ramachandran plot,* Protein Sci, 2000, Vol. 9, pp. 2322-2328.
78. **Casares S., Sadqi M., López-Mayorga O., Martìnez J.C., Conejero-Lara F.,** *Structural cooperativity in the SH3 domain studied by site-directed mutagenesis and amide hydrogen exchange,* FEBS Letter, 2003, Vol. 539, pp. 125-130.
79. **Hwang H., Pierce B., Mintseris J., Janin J., Weng Z.,** *Protein-protein docking benchmark version 3.0.,* Proteins, 2008, Vol. 73, pp. 705-709.
80. **Higurashi M., Ishida T., Kinoshita K.,** *PiSite: a database of protein interaction sites using multiple binding states in the PDB,* Nucleic Acids Res, 2009, pp. D360-364.
81. **Wennerberg K., Rossman K.L., Der C.J.,** *The Ras superfamily at a glance,* J Cell Sci , 2005, Vol. 118, pp. 843-846.
82. **Ye X., Carew T.J.,** *Small g protein signaling in neuronal plasticity and memory formation: the specific role of ras family protein,* Neuron, 2010, Vol. 68, pp. 340-361.
83. **Renault L., Kuhlmann J., Henkel A., Wittinghofer A.,** *Structural basis for guanine nucleotide exchange on Ran by the regulator of chromosome condensation (RCC1),* Cell, 2001, Vol. 105, pp. 245-255.
84. **Kent H.M., Moore M.S., Quimby B.B., Baker A.M., McCoy A.J., Murphy G.A., Corbett A.H., Stewart M.,** *Engineered mutants in the switch II loop of Ran define the contribution made by key residues to the interaction with nuclear transport factor 2 (NTF2) and the role of this interaction in nuclear protein import.,* J Mol Biol, 1999, Vol. 289, pp. 565-577.
85. **Vetter I.R., Arndt A., Kutay U., Görlich D., Wittinghofer A.,** *Structural view of the Ran-Importin beta interaction at 2.3 A resolution,* Cell, 1999, Vol. 97, pp. 635-646.
86. **Boriack-Sjodin P.A., Margarit S.M., Bar-Sagi D., Kuriyan J.,** *The structural basis of the activation of Ras by Sos,* Nature, 1998, Vol. 394, pp. 337-343.
87. **Scheidig A.J., Burmester C., Goody R.S.,** *The pre-hydrolysis state of p21(ras) in complex with GTP: new insights into the role of water molecules in the GTP hydrolysis reaction of ras-like proteins,* Structure, 1999, Vol. 7, pp. 1311-1324.
88. **Delprato A., Lambright D.G.,** *Structural basis for Rab GTPase activation by VPS9 domain exchange factors,* Nat Struct Mol Biol, 2007, Vol. 14, pp. 406-412.
89. **Eathiraj S., Pan X., Ritacco C., Lambright D.G.,** *Structural basis of family-wide Rab GTPase recognition by rabenosyn-5,* Nature, 2005, Vol. 436, pp. 415-419.
90. **Eswar N., Eramian D., Webb B., Shen M.Y., Sali A.,** *Protein structure modeling with MODELLER,* Methods Mol Biol, 2008, Vol. 426, pp. 145-159.

91. **Cavallo L., Kleinjung J., Fraternali F.,** *POPS: A fast algorithm for solvent accessible surface areas at atomic and residue level,* Nucleic Acids Res, 2003, Vol. 31, pp. 3364-3366.
92. **Kabsch W., Sander C.,** *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features,* Biopolymers, 1983, Vol. 22, pp. 2577-637.

# ACKNOLEDGMENTS

In this first part of my acknowledgments I really want to thank all the people that helped me during this PhD by a scientific point of view.

First of all my tutor Prof. Laura Bonati for the incredible devotion in her job and for all the support that she gave me during this three years, THANKS.

Fabio, for helped me in the understanding of the data mining world and for the patience to clarify all my doubts.

Then the people that compose this group and that every single day gave me support, suggestions and also lot of funny moments: Alessandro, Annalisa and Ilaria.

In addiction I really want to thank Arianna for all the helps, the suggestion and the tricks for the MD simulations and Massimiliano to support us and *lyra.*

A special thanks also to Dr. Franca Fraternali for giving me the fantastic opportunity to work in her lab in the Randall Division of the King's College. And obviously thanks also to all the people in the Randall Division that helped me during my period in London (Flavia, Pierre, Louis, Gian Felice, Stefano, Ivan, Luca, Julie).

# RINGRAZIAMENTI

In questa seconda parte dei ringraziamenti, in Italiano, invece vorrei ringraziare tutte quelle persone che per me esistono sempre e quelle persone che hanno contribuito, in un modo o nell'altro, a permettermi di fare questo percorso.

Comincio dagli amici di sempre, dalla mia famiglia, quelle persone senza le quali di certo non sarei quello che sono, grazie a tutte le esperienze che mi hanno fatto vivere insieme a loro.
La Debi, la mia figlioccia, che è una persona su cui so che potrò sempre contare, e che è riuscita a starmi vicino giorno per giorno anche quando ero a qualche chilometro di distanza, anche solo con un semplice "ciao" in chat.
Andoni, perché i viaggi fatti insieme sono qualcosa che mi porterò sempre dentro e mi hanno permesso di conoscere mezza Europa con uno zaino in spalla a tante disavventure sulla strada, ma tutte affrontate a suon di risate. (E McFly per averlo reso felice)
Causin, la Ele, la Jovi perché ognuno di voi contribuisce, ha contribuito e spero contribuirà a rendere le mie giornate quello che sono, con le vostre gioie, con i vostri casini, con tutto quello che vi riguarda.

Dopo gli amici di sempre vengono quei fantastici mattacchioni che compongono Barbie&Pistola.
Massimo " l'Enchuflatore" perché andare a mangiarsi lo gnocco fritto sotto casa mia al nostro solito tavolo non ha prezzo...
Cristian "il Proprietario", Elisa "la Fuggitiva", Martina "The Singer" perché siete stati i veri cofondatori di B&P e perché a Cavalese abbiamo vissuto delle giornate indimenticabili.
Antonio "il Cubano" e Piera "Cenerentola", perché in fondo B&P fa anche rima con Amore…
Daniele e Sonia per tutta la passione che trasmettono durante le loro lezioni.

E poi tutti gli altri che sono e restano dei fantastici compagni di ballo e non solo (Peppe, Cettina, Cristina, Agata, Adriana, Stefano, Sara, e già so che ne sto scordando una marea…)

Poi gli amici di Londra (Flavia, Pierre, Julie, Dina, Stefano, Ivan, Luca, Gian Felice, Maxime, Daniela, Simona&Simona, ed anche in questo caso sto scordando qualcuno). Perché Londra non è stata solo Randall ma è stata anche una bellissima esperienza di vita e di divertimento. Le birre al Miller, le serate al Cargo, le arrampicate… e molto altro ancora.

Ovviamente anche qui voglio ribadire tutti i miei più sentiti e sinceri ringraziamenti a tutte quelle persone che mi hanno dato una mano in quello che è stato un bellissimo percorso formativo in questi 3 anni.
Laura, per essere stata un grande capo e un punto di riferimento continuo e costante. Annalisa, Ale ed Ilaria per aver condiviso per me giorno per giorno questo periodo. Arianna per la sua immensa disponibilità e pazienza e Max, il nostro *salvatore*.
Fabio, per tutto l'appoggio datomi per permettermi di capire al meglio un mondo a me nuovo e per aver avuto la pazienza di ascoltare e chiarire tutti i miei dubbi.
Franca per avermi ospitato a Londra nel suo fantastico laboratorio e per avermi dato la possibilità di vivere l'atmosfera della Randall.

Inoltre c'è una persona che aggiungerei a questo elenco, ed è un ringraziamento che parte da molto lontano. Vorrei dire un vero e sincero grazie di cuore al Prof. Paolo Tenca del mio caro ITIS Molinari. Circa 15 anni fu la prima persona a trasmettermi la passione e l'amore per la conoscenza e fu la prima persona a scommettere sulle mie capacità. Quello è stato il mio vero slancio che mi ha portato oggi a completare questo percorso formativo.

E infine… ci sei TU… Erika… che sei riuscita a starmi accanto anche in questo folle periodo, che scommetti su di me ogni giorno, che sei li... ed eri solo da trovare ma ci sei sempre stata…

Sei entrata in modo diverso nella mia vita da poco, ma i momenti con te hanno già un enorme peso per me.

GRAZIE A TUTTI!!!!