UNIVERSITÀ DEGLI STUDI DI MILANO
FACOLTÀ DI SCIENZE MATEMATICHE FISICHE E NATURALI

Dipartimento di Chimica Organica ed Industriale
and
CNR, Istituto di Chimica del Riconoscimento Molecolare

Dottorato di Ricerca in Chimica Industriale (XXIII Ciclo)

Molecular dynamics simulations of biological macromolecules: applications to structural vaccinology and peptide design

Settore Scientifico-Disciplinare CHIM06

Tesi di Dottorato di Ricerca di:
Guido Scarabelli
Matricola R07915

Tutor: Dr. Laura Belvisi
Cotutor: Dr. Giorgio Colombo (CNR, ICRM)
Coordinatore: Prof. Dominique Roberto

Anno Accademico 2009-2010

Contents

Chapter 1

Introduction

The advent of the proteomic era, originated from the knowledge of entire genomes in the last decade, has emphasized the importance of understanding biological and functional roles of proteins present in organisms.

Proteins play a fundamental part in cell life and their damage or alteration causes the insurgence of pathologies. As a matter of fact they are important targets for drug design. In cells, proteins can be seen as molecular machines having a specific structure (called native fold) which is strictly related to their function. Often to carry it out, a protein undergoes conformational variations due to the interaction with other molecules, which can consist from small changes in a loop orientation to entire domain motions.

The interactions between proteins are important for the majority of biological functions. For example, signals from the exterior of a cell are mediated to the inside of that cell by protein-protein interactions of the signaling molecules. This process, called signal transduction, plays a fundamental role in many biological events and in many diseases (e.g. cancers). Proteins might interact for a long time to form part of a protein complex. A protein may carry another protein (for example, from cytoplasm to nucleus or vice versa in the case of the nuclear pore importins), or a protein may interact briefly with another one just to modify it (for example, a protein kinase will add a phosphate to a target protein).

Considering all the kind of chemical reactions occuring and all the different proteins present in a cell, it is very easy to understand that protein-protein interactions are of central importance for virtually every process in a living organism. Information about these interactions improves the knowledge of diseases and can provide the basis for new therapeutic approaches. For these reasons the determination of the structure, the analysis of conformational properties and the characterization of the interactions with other molecules are important parameters to study the behaviour of these biological macromolecules and to highlight how they express their function.

The relevance of computational and theoretical approaches applied to the descrip-

tion of biomolecules has been increasing over the years. Experimental techniques present difficulties in the characterization of molecular movements, for example the time scales involved are too short for any experimental measure, or it can not be easy to reproduce the conditions required. On the contrary, computational modelling studies of bio-systems provide reliable information on these structural variations which can be combined with experimental data collected in laboratories, resulting in reduction of costs and time required for discovery.

Biomolecules are probably one of the most complex subject to study considering both the high number of degrees of freedom per system (composed by macromolecules and surrounding environment) and the huge hierarchy of functionally significant timescales movements, which can vary from nanoseconds to milliseconds and beyond.

Quantum chemistry would provide the ideal knowledge to study both proteins and chemical reactions through the representation of electronic interactions and motions of light particles, such as protons, with the implementation of excellent models. Unfortunately, considering the computer power available, the number of atoms in biomolecules is too high to use it.

On the other hand methods based on classical physics can be applied to characterize macromolecules with good approximation for the analysis of such complex systems. Non-bonded interactions among molecules can be very well described by a classical potential-energy function or force field as part of a classical Hamiltonian of the system of interest, allowing a good molecule representation.

The theoretical framework of this thesis work is constituted by classical simulations applied to the analysis of peptides and proteins. Atomic and molecular degrees of freedom with the corresponding classical force fields and classical Newtonian dynamics have been used to sample the different system conformations in order to analyse and characterize their structural, dynamical and functional properties.

In particular two main topics have been studied. The first one is the charaterization and prediction of antibody binding sites on antigenic protein surfaces (epitopes). The second is the study of folding and unfolding processes of small polypeptide molecules.

1.1 Antigenic epitope prediction

The interaction between molecules is at the basis of cellular life. It is no surprise that the study of molecular recognition processes has emerged in recent years as a very important and relevant field of research, influencing areas like molecular biology, biochemistry, biomedicine, immunology, etc.

Regarding immunology, the recognition of antigens by antibodies is a fundamental step for the adaptative immune response, which leads to the neutralization of a specific pathogen. Antigens are molecules (proteins or polysaccharides) expressed on the surface of infectious agents like viruses and bacteria which are recognised by the host immune system as non-self.

In case of infection, the immune system cells identify the invading agent through specific antigenic molecules expressed on it and activate the immune response. For the organism this process is fundamental to avoid the spreading of the infectious pathogen and to neutralize it. The immune response consists at the beginning in the production of signalling molecules attracting cells like macrophages and granulocytes which start to fight the infection. Subsequently, in the activation and differentiation of limphocytes B and T which can react specifically with the production of antibodies and the activation of cells able to kill the pathogen.

When the infection is defeated, immune cells (called memory cells) remain in circle in the host body to ensure a rapid and efficient immune response through the production of specific antibodies in case of a second infection from the same pathogen. In this context, a vaccine is a biological preparation based on the capability of the immune system to keep memory of the infectious organisms. It is constituted by an inactivated pathogen, or a part of it, which is injected into an organism inducing protection against infective agents through the activation of the immune system and the production of memory cells.

In order to make vaccine preparations it is important to know which pathogen molecules can be recognised by antibodies as not all the molecules expressed on a pathogen surface can be a good target for the antibody binding. In particular, the molecular regions recognized by antibodies assume huge relevance. Being able to characterize their chemical and physical properties and determine their presence and position on molecule surfaces, can result in a considerable advantage in vaccine design, reducing both industrial cost and production time [1].

Using high resolution structures, it will be possible to design antigenic molecules making their production more efficient and improving the steps of the storing process, thereby lowering costs and eliminating problems to distribution. Furthermore with the knowledge of which parts of an antigen must be retained to preserve basic characteristics for the recognition process and which can be altered, vaccine antigens can be modified more rapidly in response to changing epidemiology. Finally, the use of

different molecules in combination-vaccines and the immunization regimens can be simplified with a proper design.

All these considerations emphasize the reasons why it is very important to identify good antigenic molecules for vaccine design. Antibody recognition is influenced by a number of different aspects related to antigenic molecules such as: size, shape, structural complexity, abundance, solubility, propensity to oligomerization, epitope structure and position and many others. Some of these aspects are in connection with the molecular structure of the antigen itself, in particular the ones related to the epitope regions. Indeed, the ability to predict epitope sequences and location is central to the development of structural vaccinology. Based on these aspects, the study of the structural and dynamical properties of these molecules, using computational techniques, can be a good instrument to improve the knowledge in this field.

The first part of this thesis work consists in the description of a new method (called MLCE, Matrix of Local Coupling Energies) developed to predict epitopes on antigenic protein surfaces. This method is based on the integrated analysis of dynamical and energetic properties of antigens in order to identify non-optimized, low-intensity energetic interaction networks in the protein structure isolated in solution. This technique relies on the idea that recognition sites may correspond to localized regions with low-intensity energetic couplings with the rest of the protein, which allows them to undergo conformational changes, to be recognized by a binding partner, and to tolerate mutations with minimal energetic expense.

Nineteen different antigenic protein structures were downloaded from the Protein Data Bank (PDB). All these proteins were crystallized alone in solution and in complex with at least one antibody. Information about the region recognised by antibodies was used only to check the predictions obtained with the method developed and not to make the predictions train. For each protein five molecular dynamics simulations of 30ns have been carried out for a total of 150ns. From the simulations the most representative structures through a cluster analysis of the conformations assumed by the proteins in the trajectories were determined. To identify epitopes, MLCE method was then applied on the main cluster structure (the most reprensetative one).

The principal cluster structure is optimized through molecular mechanics steps and on it non-bonded energy contributions are calculated for each pair of residues, resulting in a square matrix NxN (where N is the number of aminoacids) containing the energetic values. This matrix is diagonalized and decomposed in eigenvalues and eigenvectors. The first eigenvalue and its related eigenvector are subsequently used to rebuild the energy matrix multiplying it by its transpose. This process allows to filter the information in the original energetic matrix focusing only on the most important interactions present in the protein fold [2–4].

Afterwards a contact matrix (whose dimensions are still NxN) is built from the protein structure. This matrix is based on the spatial distance between each pair of residues and the matrix elements are set to 1 if two residues are close in the structure or to 0 if their distance is higher than a cut-off edge.

The energy and the contact matrices are then multiplied through the Hadamard product. The resulting matrix (called MLCE, Figure 3.3) allows to identify residues forming patches and showing coupled energetic values with the other protein regions. The patches with minimal energetic coupling correspond to sites not involved in the stabilization of the protein fold and prone to be subjected to conformational variations upon the binding of a partner like an antibody.

The contact-filtered coupling interactions are ranked in increasing order according to their respective intensities (from weaker to stronger). Starting from the minimum value (weakest local coupling interactions, defined as "soft spots", in contrast with the "hot spots", characterized by high coupling intensities), the set of putative interaction sites was defined by including increasing residue-residue coupling values until the number of couplings that correspond to the lowest 15% of all contact-filtered pairs was reached. This then is related, in the approximation, to the set of local interactions, possessing minimal intensities, which may identify antigen-antibody or protein interaction sites. The corresponding residues define putative epitope sequences.

Importantly, MLCE does not require the use of any training set of antibody-antigen complexes, as the determination of the epitope regions is based solely on structural-dynamical and energetic properties of uncomplexed antigens in isolation.

Upon analyzing the results on isolated proteins and benchmarking against antibody complexes, it has been found that the method successfully identifies binding sites located on the protein surface that are accessible to putative binding partners (Table 3.2).

To assess the predictivity performance of MLCE technique several statistical analyses on all antigens analyzed have been used. Results are reported in Table 3.2, and the average performances are similar or better than the ones reported for knowledge based methods [5].

A public web server (BEPPE, Binding Epitope Prediction from Protein Energetics) has been implemented with MLCE method. The aim is to carry out epitope predictions on single protein structures obtained from X-Ray, NMR or through computational tools like homology modelling or MD simulations. The screenshot of the main page is represented in Figure 3.9.

To use BEPPE, the input steps required are: upload a protein structure in PDB format, select the softness level of the prediction and insert an email address to receive the output.

BEPPE is implemented to allow the user to choose the prediction softness, meaning that it is possible to consider for the epitope identification the 10% (strict), 15% (default)

or 20% (soft) lowest energetic residue couplings in the MLCE calculation. As result the number of predicted residues and the patch size will vary according to the user selection (going from few residues with a strict prediction to a higher number with a soft one).

BEPPE reads the input and then performs the following passages:

- Optimize the protein structure through 500 steps of molecular mechanics
- Perform MLCE calculation
- Select the predicted residues and clusterize them into patches using the contact matrix previously calculated, rank the patches according to their average energies and select up to 4 different patches (discarding the ones composed by less than 5 residues)
- Create sequence segments in fasta format (1 residue letter code) and joining predicted residues which are far away up to 4 positions in the sequence
- Align these segments (using BLASTP algorithm [6]) with all human protein sequences available in order to find mimotopes (a mimotope is an epitope which is homologue to a region present in a host protein, making it problematic to be recognised by host antibodies [7])
- Send the output to the email address specified

The output consists in:

- predicted residues clusterized into patches
- a Pymol script readable with Pymol program, useful to have a three-dimentional representation of the protein with patches highlighted
- results of the alignment with human proteins
- a link to download output files

1.2 Folding and unfolding processes of small peptides

The second part of this work focuses the folding and unfolding processes of small polypeptides.

Protein folding is the physical phenomenon by which a polypeptide folds into its characteristic and functional three-dimensional structure from random coil.

The correct three-dimensional structure is essential for proteins to function. Indeed, failure to fold into the native shape usually produces inactive proteins with different

properties, including toxic prions, in addition to several neurodegenerative and other diseases originated from accumulation of misfolded (incorrectly folded) proteins [8]. In particular, the local formation of secondary structural elements, acting as nucleation sites for the formation of native structures, plays a crucial role in the first phases of protein folding. Thus, understanding the folding mechanism, at the highest possible resolution, of synthetic peptides adopting well-defined secondary structures is of great relevance for the comprehension of native protein folding [9, 10].

Structural studies on peptides are very important to determine their properties in order to make their use and synthesis easy and straightforward. For example, in peptide design the secondary structure is an essential element to consider (together with sequence, peptide length, hydrophobic stretches and the presence of residues like Cys or Met which are susceptible to oxidation and/or side reactions). To know the peptide structure, and how it folds in isolation, is a relevant information. For example, it is useful to avoid the formation of β sheet structures, which causes incomplete solvation of the growing peptide, resulting in a high degree of deletion sequences in the final product.

Peptides present a lot of different applications. In biomedicine they can be used as ligands for targets like receptors in order to influence and modulate their activity and all the following reaction cascade triggered in the biological response. In material science, short sequences have been utilized to characterize the sequence determinants of the formation of ordered supra-molecular structures showing nanoscale dimensions. In vaccinology, peptides can be recognised by antibodies in order to make an efficient vaccine, etc [11, 12].

The aim of this part is to identify and characterize the aminoacids mostly involved in the stabilization of a 15-mer peptide (called QK, sequence $Ac{-}KLTWQELYQLKYKGI{-}NH_2$) homologue to the region of VEGF (Vascular Endothelial Grow Factor) which interacts with its endogenous receptor and to highlight the folding/unfolding steps related to its structure. VEGF protein is involved in the angiogenesis process and constitutes a good target to regulate the formation of new blood vessels from pre-existing ones. For this reason it takes relevance in different pathologies, ranging from cardiovascular diseases to tumor and macular degeneration.

The QK polypetide, despite its small dimension, folds into an α helical shape and remains stable in this conformation even at high temperatures. Several molecular dynamics simulations were carried out on QK peptide at different temperatures in order to test the persistence of the α helix in diverse conditions. The results have been integrated with NMR spettroscopy and Circular Dichroism (CD) data obtained in collaboration with Dr L. D'Andrea's group at IBB-CNR.

After the analysis of the conformations assumed by the peptide during the MD simulations, all data collected agreed in identifying two leucine residues (in position

7 and 10) as fundamental for the stability of the structure. The time spent in α helix by the peptide decreased as the temperature rised, but even at the highest temperature simulated (380K) have been found helical conformations. This fact was confirmed by the experimental analysis. In addition, the folding pathway was studied with a simulation started from the extended peptide conformation. Results showed that after 50ns of MD the peptide was folded in α helical structure.

To further improve the knowledge on this peptide a mutation on its sequence has been induced, in order to substitute the leucine in position 10 with an alanine residue, and check if the stability of the helix could be perturbed. With this mutation the polypeptide was less stable at higher temperatures and it unfolded with higher frequency during the simulations. Again, these results were in agreement with the NMR and CD analysis, demonstrating the importance of the formation of a hydrophobic interaction between leucine residues in position 7 and 10 for the stability of the structure.

Overall, results obtained proved to be relevant for the determination of epitope antigenic regions on protein surfaces and for the characterization of folding/unfolding processes of small peptides. In both cases, despite the different areas studied, we demonstrated the importance of the analysis of molecular properties with computational techinques and their relevance to support experimental research.

Publications on which this thesis is based

- **Scarabelli G**, Morra G, Colombo G. **Predicting interaction sites from the energetics of isolated proteins: a new approach to epitope mapping.** Biophys J. 2010 May 19;98(9):1966-75.

- Soriani M, Petit P, Grifantini R, Petracca R, Gancitano G, Frigimelica E, Nardelli F, Garcia C, Spinelli S, **Scarabelli G**, Fiorucci S, Affentranger R, Ferrer-Navarro M, Zacharias M, Colombo G, Vuillard L, Daura X, Grandi G. **Exploiting antigenic diversity for vaccine design: the Chlamydia ArtJ paradigm.** J Biol Chem. 2010 Sep 24;285(39):30126-38. Epub 2010 Jun 30.

- Diana D, Ziaco B, Colombo G, **Scarabelli G**, Romanelli A, Pedone C, Fattorusso R, D'Andrea LD. **Structural determinants of the unusual helix stability of a de novo engineered vascular endothelial growth factor (VEGF) mimicking peptide.** Chemistry. 2008;14(14):4164-6.

- Diana D, Ziaco B, **Scarabelli G**, Pedone C, Colombo G, D'Andrea LD, Fattorusso R. **Structural analysis of a helical peptide unfolding pathway.** Chemistry. 2010 May 10;16(18):5400-7.

Chapter 2

Methods of simulations

Because of the impossibility to observe individual atoms or molecules directly, various models have been developed to describe and characterize the molecular properties of a system. In this respect the personal image of an atom or molecule is strongly dependent on the models that have been chosen to represent them. The sophistication of the model is related to the properties that will be analysed, and in general it is advisable to choose the simplest representation that considers the properties of interest in a satisfactorily way.

Over the last decade a combination of computer graphics and molecular modelling techniques has resulted in unprecedented power to create and manipulate three dimensional models of molecules using computers. The aim of this chapter is to provide a basic description of all-atom molecular dynamics (MD) simulations of peptides and proteins and of the analysis of the resulting sets of data.

Molecules such as proteins, lipids, DNA, carbohydrates are dynamic and their parts undergo movements which can lead to changes in molecule conformations. Often experimental analyses measure a time average or an ensemble average over the range of possible configurations the molecule can adopt. One way to investigate the range of accessible configurations is to simulate the motions or dynamics of a molecule numerically. This can be done by computing a trajectory, i.e. a series of molecular configurations as a function of time by the simultaneous integration of the Newton's equations of motion.

MD simulations are based on the time dependent behaviour of atomic and molecular systems calculation, giving a detailed description of the variation from one conformation to another of the system studied. Simulations generate ensembles of representative configurations in such a way that accurate values of thermodynamic and structural properties can be obtained with a reasonable amount of computation, in particular statistical analysis links microscopical and macroscopical properties providing the fundamental principles for the description of biomolecular systems.

In the first of part of the chapter, calculation of time and ensemble average properties of molecular systems are described, subsequently, concepts and equations at the basis of Molecular Dynamics (MD) simulations are explained. In the second part are highlighted in detail the determination of thermodynamic and structural properties deriving from a trajectory.

2.1 Time and Ensemble Averages

Macroscopic properties like pressure, heat capacity, volume etc depend on the positions and momenta of the N particles constituting the system. The value of a particular property A at a certain time t can be defined as a function of $\mathbf{p}^N(t)$ and $\mathbf{r}^N(t)$ representing the N momenta and positions of the particles at time t, respectively. The instantaneous value of A at time t can thus be written as:

$$A(\mathbf{p}^N(t), \mathbf{r}^N(t)) = A(p_{1x}, p_{1y}, p_{1z}, p_{2x}, p_{2y}, p_{2z}, \ldots, x_{1x}, x_{1y}, x_{1z}, x_{2x}, x_{2y}, x_{2z}, \ldots, t) \tag{2.1}$$

where p_{1x} corresponds to the momentum of particle 1 in the x direction and x_{1x} is its x coordinate. During time, the value of quantity A changes because of the effect of temperature fluctuations and interactions between particles.

Experimentally it is impossible to measure the single value of A at time t, but it is possible to measure the average of A during the time in which the experiment is carried out, and therefore it represents a time average. As the time over which the measurement is made grows to infinity, the average value of A approaches to its real equilibrium value. The average value of A_{ave} can thus be written as:

$$A_{ave} = \lim_{\tau \to \infty} \frac{1}{\tau} \int_{t=0}^{\tau} A(\mathbf{p}^N(t), \mathbf{r}^N(t)) dt \tag{2.2}$$

In this case, the measurement time is much longer than the typical relaxation time of each event and the average value represents the equilibrium one.

In theory moving to simulations for the calculation of average values of selected system properties is straightforward: starting from an energy function which describes the interactions between the particles in the system it is possible to calculate the forces acting on the particles, then through Newton's second law of motion the positions and momenta as a function of time are obtained.

Application of equation 2.2 would then provide the average values of the property of interest, but unfortunately the dimensions of real molecular systems make it impossible to calculate. The determination of interactions, positions and momenta for number of particles of the order of 10^{23} is currently out of reach for most powerful computers even

using very simplified energy functions.

This problem can be overcame with statistical mechanics. In statistical mechanics the attention is not focused just on one single system evolving in time, but rather on a large number of replicas of the same system evolving simultaneously. As a consequence the time average is replaced by an ensemble average

$$< A >= \int \int d\mathbf{p}^N d\mathbf{r}^N A(\mathbf{p}^N, \mathbf{r}^N) \rho(\mathbf{p}^N, \mathbf{r}^N) \tag{2.3}$$

$< A >$ corresponds to the ensemble average or expectation value of property A, i.e. the average value of A over all the replicas of the system in the ensemble generated by the simulation. $\rho(\mathbf{p}^N, \mathbf{r}^N)$ is the probability density of the ensemble, meaning the probability to obtain a configuration with momenta $\mathbf{p}^N$ and positions $\mathbf{r}^N$ among all the configurations sampled in the simulation. If the simulation is long enough to sample all the relevant configurations for the system for the ergodic hypothesis the ensemble average will be equivalent to the time average. Under these conditions the density of probability is described by the typical Boltzmann distribution.

$$\rho(\mathbf{p}^N, \mathbf{r}^N) = \frac{1}{Q} e^{\frac{-E(\mathbf{p}^N, \mathbf{r}^N)}{k_B T}} \tag{2.4}$$

where E is the energy function, Q the partition function, k_B the Boltzmann's constant and T the temperature. The partition function is generally written in terms of the Hamiltonian H governing the system, e.g.

$$Q_{NVT} = \frac{1}{N!} \frac{1}{h^{3N}} \int \int d\mathbf{p}^N d\mathbf{r}^N e^{\frac{-H(\mathbf{p}^N, \mathbf{r}^N)}{k_B T}} \tag{2.5}$$

The subscript NVT indicates a systems with a constant volume V, number of particles N and temperature T (Canonical Ensemble). In MD simulations on biological systems the Hamiltonian H can be approximately considered equal to the total energy E of the system. N! arises from the indistinguishability of the particles in a system to ensure proper counting of states, while $\frac{1}{h^{3N}}$ is related to the equivalence of the partition function to that calculated through quantum mechanics.

MD simulations generate a trajectory consisting of a collection of subsequent configurations and describing how the dynamic variables vary in the time. Thermodynamic quantities are calculated from the trajectory using numerical integration of equation 2.3:

$$< A >= \frac{1}{M} \sum_{i=1}^{M} A(\mathbf{p}^N, \mathbf{r}^N) \tag{2.6}$$

where M is the number of configurations (samples) from the simulation over which the property is evaluated.

2.2 Molecular Dynamics

Molecular Dynamics (MD) is a technique useful to compute the equilibrium and transport properties of many-body systems using classical mechanics laws. It constitutes an excellent approximation for a wide range of materials and applications for systems where electronic motions and reorganizations are not involved, like for the dynamical characterization of complex biomolecular molecules or a polymer behavior. The final result of an MD simulation is a trajectory that highlights the variation of the positions and velocities of the atoms in time. On this trajectory it is possible to determine the properties of interest as time or ensemble averages, as described previously. The successive configurations composing the trajectory of the system are generated through the application of Newton's laws of motion on atom particles:

$$\frac{d^2 x_i}{dt^2} = \frac{F_{xi}}{m_i} \tag{2.7}$$

From the equation 2.7 it is evident that particle i with mass m_i moves along coordinate x_i subjected to the force F_{xi} which is due to the presence and interaction of atom i with all other particles in the system. The force can be expressed as the negative derivative of a potential function $V(r_1, r_2, r_3, \ldots, r_N)$ describing the fundamental types of interactions in the system.

$$F_i = -\frac{\partial V}{\partial r_i} \tag{2.8}$$

V can be considered as the potential energy of the system as a function of atomic positions, the equations are solved simultaneously in small time steps (dt) and the atomic coordinates are written in the trajectory output file.

A typical potential function for all-atom protein simulations is expressed in the following form:

$$V(R) = \sum_b D_b[1 - e^{(-a(b-b_0))}]^2 + \sum_\theta H_0(\theta - \theta_0)^2 + \sum_\phi H_\phi[1 + scos(n\phi)] + \sum_\chi H_\chi \chi^2 +$$
$$+ \sum_b \sum_{b'} F_{bb'}(b - b_0)(b' - b'_0) + \sum_\theta \sum_{\theta'} F_{\theta\theta'}(\theta - \theta_0)(\theta' - \theta'_0) +$$
$$+ \sum_b \sum_\theta F_{b\theta}(b - b_0)(\theta - \theta_0) + \sum_\theta \sum_{\theta'} F_{\theta\theta'\phi}(\theta - \theta_0)(\theta' - \theta'_0)cos\phi +$$
$$+ \sum_\chi \sum_{\chi'} F_{\chi\chi'} \chi\chi' + \sum_i \sum_{j>i} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} + \frac{q_i q_j}{r_{ij}} \right] \tag{2.9}$$

Each term defines the contribution of a particular element:

1. $\sum_b D_b[1 - e^{(-a(b-b_0))}]^2$ represents the stretch of a chemical bond
2. $\sum_\theta H_0(\theta - \theta_0)^2$ corresponds to the deformation of a bond angle
3. $\sum_\phi H_\phi[1 + scos(n\phi)]$ reflects the torsion of a dihedral angle
4. $\sum_\chi H_\chi \chi^2$ takes into account the distorsion of atom involved in planar bonds laying outside the plane
5. $\sum_b \sum_{b'} F_{bb'}(b - b_0)(b' - b'_0)$ defines the contemporary stretching distorsion of two bonds
6. $\sum_\theta \sum_{\theta'} F_{\theta\theta'}(\theta - \theta_0)(\theta' - \theta'_0)$ represents the contemporary bend of two angle bonds
7. $\sum_b \sum_\theta F_{b\theta}(b - b_0)(\theta - \theta_0)$ corresponds to the stretch of one bond and the bend of an angle bond
8. $\sum_\theta \sum_{\theta'} F_{\theta\theta'\phi}(\theta - \theta_0)(\theta' - \theta'_0)cos\phi$ defines the distorsion of an angle bond and a dihedral one
9. $\sum_\chi \sum_{\chi'} F_{\chi\chi'} \chi\chi'$ reflects the stretch-bend deformation of atoms forming planar bonds
10. $\sum_i \sum_{j>i} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} + \frac{q_i q_j}{r_{ij}} \right]$ considers non-bonded interactions (Van der Waals and electrostatic).

Several parameters are present in the Force Field, such as the equilibrium distances, force constants or Van der Waals and electrostatic terms, all of them have to be determined by either using experimental data or through the fitting to high level ab initio calculations. Clearly, the parametric nature of the force fields imposes restrictions to their uses in contexts which are different from the ones they have been developed for, so for example it would be unwise to use a force field set for aminoacids in a inorganic polymer study.

2.2.1 Molecular Dynamics Parameters

First, to run an MD simulation it is necessary to define a molecular system to study and to identify the properties useful for its characterization. Generally, the model system will consist of N particles which will interact under the action of the potential and forces defined in equations 2.7 to 2.9.

An MD trajectory is divided into two parts, the first one is the equilibration stage, in which the system (and the properties of interest) will evolve as a function of time, and the second one is the production phase where it is possible to carry out the effective measurements as the system has reached the equilibrium. The choices of the model, the equilibration time and the way the measurement is carried out are very sensitive points which have to be evaluated carefully. Indeed incorrect results or bad artifacts can be generated by using the wrong model to describe the phenomena, by using too short equilibration/measurement times, or by not noticing irreversible and chemically meaningless changes that can occur in the system.

Starting conformation

To start the simulations initial positions and velocities to all atoms in the system must be assigned. In the case of biological molecules or protein simulations, the initial positions can be obtained from structural determination experiments such as NMR or X-Ray measurements.

Clearly, in biomolecular simulations, the user is mostly interested in investigating the properties of the system in presence of the appropriate solvent (or mixture of solvents), rather than simply studying gas-phase properties. To this end, the solute (protein, DNA, drugs, etc. . .) is inserted in a pre-equilibrated solvent bath (any solvent molecules whose coordinates are too close to the solute atoms are eliminated from the system). In theory a simulation should be able to reproduce the behaviour of an infinite system or of a real system of around 10^{23} particles, in which a negligible number of particles would be in contact with the boundaries (like the vessel walls in real-life experiments), in order to calculate straightforwardly macroscopic quantities. In practice this situation is completely out of reach even for the most powerful computers, and the study has to be carried out on finite-size systems characterized by some boundaries.

The correct choice of the method to treat the boundaries of the simulation is fundamental for the calculation of the properties of interest. Depending on the physico-chemical problem under investigation it is possible to set up two different boundary conditions, Periodic Boundary Conditions and Non-Periodic Boundary Methods. In this work, we will only use Periodic Boundary Conditions (PBC) and explicit solvent representations which are the most used for protein and peptide studies.

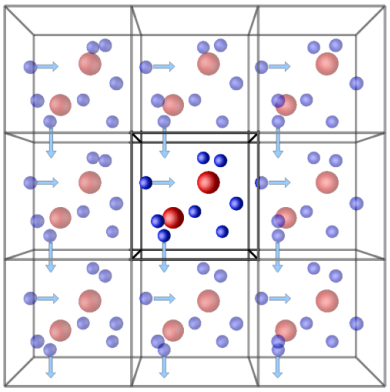


Figure 2.1: *Example of a cubic simulation cell surrounded by its replicas (8 in this case) in order to set up periodic boundary conditions in 2D.*

Periodic Boundary Conditions

Periodic Boundary Conditions (PBC) are useful to run a simulation considering a relatively small number of particles, in such a way that the particles experience interactions and forces as if they were in a bulk fluid. The simplest representation of such a system is represented by a cubic box of particles which is replicated in all directions to give a periodic array (Figure 2.1).

The particles coordinates in the replica images are obtained by adding to the original ones multiples of the box sides. If a particle leaves the box during the simulation, it will be replaced by its image coming in from the opposite side of the box. In this way the number of particles in the simulation box is kept costant and the solvent behaves basically as a bulk with no border effects affecting the results of the simulation. To reduce the cost in term of calculation periodic boundary conditions are most often combined with the minimum image convention: only one, the nearest-image of each particle is considered for short-range non-bonded interaction terms.

There is also the possibility of choosing different shapes of the box, the most common are types of geometric solid (cubic, hexagonal prism, truncated octahedron, rhombic dodecahedron and elongated dodecahedron) that fulfil the condition to fill the whole space and be replicated by simple translation operations of the central box in Cartesian coordinates, in order to avoid problems with the periodic boundary conditions.

The choice of the type of box depends on the shape and simmetry of the molecules under study, for example for globular molecules with an approximately spherical symmetry the best one will be the truncated octahedron or the rhombic dodecahedron, which can approximate better the spherical shape. For extremely elongated systems with a cylindrical symmetry, the best choice will be the hexagonal prism, while for systems like membrane double layers, a good box is a cubic or parallelepiped one.

PBC conditions are very convenient and are as close as possible to a realistic situation even though they present some drawbacks. For example it is not possible to

achieve fluctuations that have a bigger wavelength than the length of the cell itself. Also if there are long-range interactions between molecules, such as electrostatic ones, it is often necessary to use truncation or cutoff approximations for the treatment of long range forces in order to avoid the formation of artifacts such the ones where a particle interacts with its own periodic image (and so with itself). For these reasons it is very important to set up proper conditions in order to have a correct simulation box.

After the solvation of the system initial atomic positions must be carefully checked to avoid any sizeable overlap of groups and different atoms positioned in the same point in the space box. On average initial structures have to be minimized in order to remove bad contacts and optimize bond, angular and torsional interactions. The minimization procedure has the main objective to place the initial structure on low energy points on the Potential Energy Surface. For the starting movement, to each particle is assigned a initial velocity that is drawn from a uniform Maxwellian distribution of velocities consistent with the temperature at which the simulation will be run.

$$p(\mathbf{v}_i) = \sqrt{\frac{m_i}{2\pi kT}} e^{\frac{m_i \mathbf{v}_i^2}{2kT}} \tag{2.10}$$

Electrostatic interactions

The most time consuming part in an MD simulations is the calculation of forces. For example considering the potential describing the interactions in the system like the one depicted in equation 2.9 (which is essentially a pair-wise potential) it is necessary to take into account the force contribution acting on particle i due to the presence of all its neighbors. This would imply that the computational time required to evaluate the forces is in the range of N^2, where N is equal to the number of atoms in the system.

The biggest part of the computational cost is related to the term representing the non-bonded interactions energies (Lennard-Jones and Electrostatics) in equation 2.9. In principle, these interactions are calculated between every pair of atoms in the system, but actually for many systems this is not strictly necessary. Lennard-Jones potentials (the C_6 and C_{12} terms) for instance decay very fast as a function of the distance between atoms, reflecting the r^{-6} dependence of the dispersion interactions. The most simple way to deal with them is to use a non-bonded cutoff and to apply the minimum-image convention. With a cutoff threshold the interactions between all pairs of atoms whose distance is higher than the cutoff are set to zero and excluded from the calculation.

When periodic boundary conditions are used, the box dimensions should be large enough to avoid that a particle can interact with their own periodic images (which leads to problematic artifacts that invalidate the simulation). This limits the extension of the cutoff to be no more than half the length of the cell when using a cubic cell.

Cutoff values should be chosen carefully when long range electrostatic interactions are involved in the simulations, such as in the case of presence of ionizable groups, counterions and so on. Of course in these kind of situations the thresholds should be much longer than in the cases of uncharged Lennard-Jones particles.

In general the use of cutoffs leads to truncating the potentials in non natural way and it brings errors especially in the cases of charged systems, where electrostatics is particularly important. To solve these problems other methods have been developed to properly deal with long-range interactions. In the next section it will be described the treatment of long-range electrostatic interactions with Particle Mesh Ewald (PME) method, a technique that allows to treat explicitly electrostatics without the use of cutoff thresholds.

Ewald summation method

The aim of Ewald summation method is to model properly long-range forces in simulations involving charged groups [13, 14]. Electrostatic forces are the main group of long-range interactions and decay as r^{-1} and their treatment is fundamental for the calculation of properties such as the dielectric constant.

This method has been developed to compute long-range contributions to the potential energy in systems with periodic boundary conditions derived from ionic crystal studies. In a system positively and negatively charged particles are assumed to be located in a cube with side of length L (Volume = L^3), with periodic boundary conditions, and the total number of particles in the simulation box is equal to N (of course at short distances the particles repel one another). The system is electrically neutral, meaning that ($\sum_i z_i = 0$ with z_i corresponding to the charge of the atom i).

Coulomb contribution to the potential energy of the system is equal to:

$$U_{Coul} = \frac{1}{2} \sum_{i=1}^{N} z_i \phi(r_i) \tag{2.11}$$

with $\phi(r_i)$ is the electrostatic potential at the position of ion i:

$$\phi(r_i) = \sum_{j,n}^{'} \frac{z_j}{|r_{ij} + nL|} \tag{2.12}$$

Where the prime sign on the summation in equation 2.12 indicates that the sum is calculated for all periodic images n and over all particles j, except j=i if n=0.

As the use of 2.12 to study electrostatic properties in a simulation would result in a very slowly converging sum, the expression for charge density have to be rewritten.

In equation 2.12 the charge density has been described as a sum of d functions, and the contribution to the electrostatic potential due to these charges decays as r^{-1}. Taking as example a situation in which every charge z_i is surrounded by a diffuse charge distribution of the opposite sign, such that the total charge of the cloud exactly cancels z_i, the electrostatic potential due to particle i is due exclusively to the fraction of z_i which is not screened. At long distances, this fraction rapidly goes to zero influenced by the functional form of the screening charge distribution. Generally, the choice for this distribution is a Gaussian.

The contribution to the electrostatic potential at a point r_i due to a set of screened charges can be easily computed by direct summation, because the electrostatic potential due to a screened charge is a rapidly decaying function of r. However, in MD simulation it is not relevant the potential due to the screened charge, but the one due to actual point charges. Hence, it must be corrected for the fact that it has been added a screening charge cloud to every particle.

This compensating charge distribution is a function that varies smoothly in space, for example to compute the electrostatic energy at the site of ion i it is necessary firstly to exclude the contribution of the charge z_i to the electrostatic potential. However, it is convenient to add a screening charge around ion i to the compensating charge distribution that must be subtracted as the compensating charge distribution becomes a smoothly varying periodic function. The idea is that such a function can be represented by a rapidly converging Fourier series. Of course, at the end it is necessary to correct for the inclusion of a spurious self interaction between ion i and its compensating charge cloud.

In this way PME has the advantage of a convergence of the Fourier-space summation compared to its real-space equivalent when the real-space interactions are long-ranged. Because electrostatic energies consist of both short- and long-range interactions, it is maximally efficient to decompose the interaction potential into a short-range component summed in real space and a long-range component summed in Fourier space.

Integration of equation of motion

MD simulations are based on the integration of Newton's equation of motion. To calculate the position of a particle at a certain time $t + \Delta t, r(t + \Delta t)$ it is possible to expand with Taylor series the coordinate of that particle around time t.

$$r(t + \Delta t) = r(t) + v(t)\Delta t + \frac{f(t)}{2m}\Delta t^2 + \frac{\Delta t^3}{3!}r + O(\Delta t^4) \tag{2.13}$$

and in the same way:

$$r(t - \Delta t) = r(t) - v(t)\Delta t + \frac{f(t)}{2m}\Delta t^2 - \frac{\Delta t^3}{3!}r + O(\Delta t^4) \tag{2.14}$$

Summing up equations 2.13 and 2.14 results in:

$$r(t + \Delta t) + r(t - \Delta t) = 2r(t)\frac{f(t)}{m}\Delta t^2 + O(\Delta t^4) \tag{2.15}$$

and then:

$$r(t + \Delta t) \approx 2r(t) - r(t - \Delta t) + \frac{f(t)}{m}\Delta t^2 \tag{2.16}$$

The estimate of the new positions contains an error which is of the order of Δt^4, where Δt is the time step in out MD scheme, typically between 1 and 10 fs in the simulation of biomolecular systems. The algorithm presented here is the so called Verlet algorithm. It is important to notice from 2.16 that velocities are not used to compute the new position, however, it is always possible to compute the velocity from the knowledge of the trajectory using:

$$r(t + \Delta t) - r(t - \Delta t) = 2v(t)\Delta t + O(\Delta t^2) \tag{2.17}$$

or

$$v(t) = \frac{r(t + \Delta t) - r(t - \Delta t)}{2\Delta t} + O(\Delta t^2) \tag{2.18}$$

This expression for the velocity is exact to within an order Δt^2, but it is not possible to obtain a better estimate of the velocities (and hence of the kinetic energy and temperature of the system) using Verlet-like algorithms.

Once the new positions have been calculated the old ones (at $t - \Delta t$) are discarded. After each time step, it is possible to calculate current temperature, potential energy and total energy. Of course the total energy must be conserved, and a good integration scheme has to ensure this in the very first place.

It is possible to use other integration algorithms, for example, GROMACS relays on leap-frog algorithm for the integration of the equations of motion. This algorithm uses positions r at time t and velocities v at time $t - \Delta t/2$ and it updates the positions and velocities using the force f(t) determined by the positions at time t:

$$v\left(t + \frac{\Delta t}{2}\right) = v\left(t - \frac{\Delta t}{2}\right) + \frac{f(t)}{m}\Delta t \tag{2.19}$$

and

$$r(t + \Delta t) = r(t) + v\left(t + \frac{\Delta t}{2}\right)\Delta t \tag{2.20}$$

Temperature and pressure control

To obtain an easier connection with experiments it is often desirable to run simulations at constant Temperature (T) or Pressure (P), the two most common simulation ensembles are in fact the NVT and NPT. This section is about the methods used to control the temperature and the pressure in MD simulations.

The temperature regulation in MD simulations is fundamental for comparison of results with experimental findings. Moreover it might be of interest to investigate the behavior of a system at different temperatures or to check the temperature induced unfolding behavior of a protein, DNA stretch etc. Simulated annealing MD protocols, in which the temperature is changed in a controlled fashion, are also of interest for NMR or X-ray structure refinements.

Temperature is related to the average kinetic energy:

$$< K >_{NVT} = \frac{3}{2} N K_B T \tag{2.21}$$

From this relation is evident that it is easily possible to control the temperature scaling the kinetic energy and hence the atom velocities. An alternative way to maintain the temperature close to the desired value is to couple the system to an external heat bath kept at the desired temperature. In the Berendsen coupling algorithm the bath acts as a heat reservoir which can supply or remove energy from the system. The velocities are scaled at each time step, such that the rate at which the temperature changes is proportional to the difference in temperature between the bath and the system. If T_0 is defined as the reference temperature and T as the instantaneous one the scaling formula will be:

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau} \tag{2.22}$$

meaning that a temperature deviation decays exponentially with a time constant τ.

This method presents the advantage related to the fact that the strength of coupling can be varied and adapted to different situations by just changing the scaling factor τ. For equilibration purposes for instance the coupling time can be taken quite short (e.g. $0.01ps$), while for productive runs this value can be much higher (like $0.5ps$), in general if τ is high the coupling is weak and viceversa (when it equals the time step, we simply have a velocity rescaling algorithm).

The main problem with this algorithm is that it does not generate rigorous canonical averages, velocities are rescaled artificially and this is reflected in any temperature difference between components of the system (this is particular relevant in the phenomenon of "hot" solvent and "cold" solute). One possible solution is to couple separately the different components to the heat bath, but a problem of unequal energy distribution among components may still be present.

Just like the temperature in a simulation it is also possible to control the pressure. This is useful to check the system behavior as a function of pressure, enabling the study of conformational changes induced by ultra-high pressure conditions. These types of studies are being applied for instance to study enzymatic reactivity for industrial applications or in the study of the properties of proteins which show conformational variations.

Pressure fluctuations are generally much more pronounced than temperature ones since the pressure is related to the virial, which is obtained as the product of the positions and the derivative of the potential energy function. This product is much more sensitive to the variations in position than the internal energy, which brings bigger pressure fluctuations.

Berendsen control algorithm can be used also for the pressure control, in fact it is based on the same concepts as for the temperature regulation. This algorithm rescales coordinates and box vectors every step with a matrix μ, which has the effect of a first-order relaxation of the pressure towards a given reference pressure P_0.

$$\frac{dP}{dt} = \frac{P_0 - P}{\tau_P} \tag{2.23}$$

and the scaling matrix μ is:

$$\mu_{ij} = \delta_{ij} - \frac{\Delta t}{3\tau_P}\beta_{ij}[P_{0ij} - P_{ij}(t)] \tag{2.24}$$

where β is the isothermal compressibility of the system. In most cases this will be a diagonal matrix, with equal elements on the diagonal, the value of which is on average not known. Generally it is sufficient to consider a rough estimate of the value of β

since it only influences the non-critical time constant of the pressure relaxation without affecting the average pressure itself. For water at 1atm and 300K this value is equal to $4.6 * 10^{-5} Bar^{-1}$. The scaling can be done isotropically or anisotropically depending on the type of system being simulated. For instance for a globular protein isotropic scaling is fine while for systems with interfaces (like membrane receptors) or with conformational changes in one direction mainly anisotropic scaling is better.

2.2.2 Thermodynamic quantities

Computer simulations allow the calculation of quantities which can be compared directly with experimental results (a fundamental step for the validation of simulation results) and also the prediction of properties inaccessible to experiments. Many different types of properties can be calculated ranging from average energies to structural and conformational information, in the next part of the paragraph some of them will be explained.

Energy

Internal energy is the most straightforward quantity that it is possible to calculate from simulations. It can be obtained as the ensemble average of the energies of the states (configurations) sampled during simulations:

$$< E >= \frac{1}{M} \sum_{i=1}^{M} E(\mathbf{p}^N, \mathbf{r}^N) \tag{2.25}$$

Pressure

In simulations generally pressure is determined through the use of the virial theorem. The virial is defined as the expectation value of the sum of products of the particle coordinates with the forces acting on them. This can be formalized as:

$$W = \sum x_i p'_{xi} \tag{2.26}$$

with x_i being the coordinate and p'_{xi} being the first derivative of the momentum along that coordinate. The latter quantity has the dimensions of a force. The virial theorem states that the virial is equal to $-3Nk_BT$. By knowing from the simulation the forces acting on each single atom it is possible to obtain the pressure through this expression:

$$P = \frac{1}{V}\left[Nk_BT - \frac{1}{3K_BT}\sum_{i=1}^{N}\sum_{j=i+1}^{N} r_{ij}f_{ij}\right] \tag{2.27}$$

Temperature

The system temperature can be calculated directly from the kinetic energies K of the particles:

$$K = \sum_{i=1}^{N}\frac{|p_i|^2}{2m_i} = \frac{k_BT}{2}(3N - N_C) \tag{2.28}$$

The single atom masses m_i, their momenta etc. are all known from the trajectories of MD simulations and N_c takes into account the removal of 3 degrees of freedom to constrain the total momentum of the system to a constant equal to zero.

2.2.3 Structural properties

Molecular simulations give access to the calculation of a large number of structural properties, in special they are relevant for the study of conformational variations, molecular interactions, in the description of allosteric modifications, etc.

Root Mean Square Deviation

Root Mean Square Deviation (RMSD) generally contains information on the divergence in time of a structure from a reference one and it is determined with the following formula:

$$RMSD(t_1, t_2) = \sqrt{\frac{1}{M}\sum_{i=1}^{N} m_i \|r_i(t_1) - r_i(t_2)\|^2} \tag{2.29}$$

Where M is the sum over all the atom masses, and $r_i(t)$ is the position of atom i at time t. A protein is usually fitted on the backbone atoms N, C_α, C or just C_α atoms, but of course it is possible to compute the RMSD over other elements like only side chain atoms or even the whole protein. In general as reference structure for the calculation it is used the first one in the simulation (or a crystal one). In addition it can be defined a matrix with the RMSD as a function of t_1 and t_2, allowing easy identification of structural transitions in a trajectory.

Radius of gyration

The Radius of Gyration (Rg) gives a measure for the compactness of the structure and it is defined as:

$$R_g = \sqrt{\frac{\sum_i ||r_i||^2 m_i}{\sum_i m_i}} \tag{2.30}$$

This measure is very useful in the polymer field in order to describe the dimensions of a polymer chain, in MD simulations it gives information about the changes in the protein structure shape.

Root Mean Square Fluctuation

Root Mean Square Fluctuation (RMSF) determines the average fluctuation of each protein atom during the simulation time using as reference structure the average one:

$$RMSF = \sqrt{\frac{1}{T}\sum_{t_j=1}^{T} ||r_i(t_j) - \tilde{r}_i||^2} \tag{2.31}$$

where T is the time used to determine the average and $\tilde{r}_i$ is the average position of particle i.

Cluster analysis

Molecular simulations generate a large amount of data representing different conformations of the molecules studied. Many of these conformations are very similar and in these cases it is desirable to filter out from the large data set only a subset of representative conformations that can be much more easy to analyse avoiding redundance. This is carried out generally by using statistical cluster analysis methods, which can group together objects with a certain similarity and extracting the simulation representative structures. For this process it is necessary to define a measure for similarity between objects; generally in protein and peptide simulations a natural choice is the RMSD. However it does not exist a universal similarity measure nor a general cluster analysis method to be applied to every situation without a previous assessment of the problem at hand.

In this paragraph it is briefly described the method introduced by Daura and coworkers [15] which is actually based on recursive RMSD evaluations: to find clusters of structures in a trajectory the RMSD of atom positions between all pairs of atoms is determined. For each structure the number of other structures for which the RMSD is

less than a threshold value (neighbour conformations) is calculated. Then the structure with the highest number of neighbours is taken as the center of the cluster and forms, together with its neighbours, the first cluster. Subsequently the structures of this cluster are thereafter eliminated from the pool of structures. The process is repeated until the pool of structures is empty. This procedure generates a series of nonoverlapping clusters of structures.

Secondary structure analysis

Another useful parameter to analyze protein structures is the evolution of secondary structure. The DSSP (Dictionary of Secondary Structure of Proteins) is the most widely used approach in this respect. The DSSP program was designed by Wolfgang Kabsch and Chris Sander to standardize secondary structure assignment [16]. DSSP is a database of secondary structure assignments for all protein entries in the Protein Data Bank (PDB) and also the program that calculates DSSP entries from PDB entries.

Normal Mode analysis

Normal mode analysis (NMA) leads to the expression of protein dynamics in terms of a superposition of collective variables, namely, the normal mode coordinates [17].

The main idea is that large and relevant protein movements may actually be under selective pressure. For this reason, aminoacid sequences may have evolved so that energetic barriers related to big structural variations (necessary to carry out the protein function) are low, as consequence, analysing one or a few low-frequency normal modes of the protein can be a fast and reliable method to describe them.

NMA is most often used in order to try to guess what kind of conformational change a protein undergoes in order to fulfil its function, by analysing its lowest-frequency modes one after the other.

Normal mode calculation is based on the harmonic approximation of the potential energy function around a minimum energy conformation. This approximation allows the analytic solution of the equations of motion by diagonalizing the Hessian matrix (the mass-weighted second derivatives of the potential energy matrix). The eigenvectors of this matrix are the normal modes, and the eigenvalues are the squares of the associated frequencies. The protein movement can be represented as a superposition of normal modes, fluctuating around a minimum energy conformation. For proteins, the normal modes responsible for most of the amplitude of the atomic displacement are associated to the lowest frequencies. In order to avoid time-consuming energy minimizations, as well as the corresponding drift of the studied structure, a single-parameter Hookean potential is used, which was shown to yield low-frequency normal modes as accurate as those obtained with more detailed, empirical, force fields [18].

$$E_P = \sum_{d^0_{i,j<R_c}} c(d_{i,j} - d^0_{i,j})^2 \tag{2.32}$$

where $d_{i,j}$ is the distance between two atoms i and j, $d^0_{i,j}$ is the distance between the atoms in the three-dimensional structure, c is the spring constant of the Hookean potential (assumed to be the same for all interacting pairs) and R_C is an arbitrary cut-off, beyond which interactions are not taken into account. This approximation implies that the reference structure represents the minimum energy conformation.

Moreover, all atom masses are set to the same fixed value in the kinetic energy term, as this approximation was shown to have little influence on the low-frequency modes. Therefore, only normalized frequencies are reported, the lowest non-trivial frequency being set to one. Note that there are always six zero frequencies (corresponding to the three overall rotations and three overall translations of the system), but more than six can be obtained if a group of atoms is at a distance larger than R_C from the others.

Major applications of normal modes are the identification of potential conformational changes like in enzyme structures upon ligand binding, membrane channel opening or after generic protein-protein interaction events.

Energy Decomposition Method

A protein structure is composed by an intricated network of bonded and non-bonded chemical interactions. All the bonds that join the protein atoms of the backbone are built during the traslation process in the ribosomes, but it is important to underline that to assume the folded and functional state non-bonded interactions are fundamental. Indeed, a protein structure can be seen as a complex network of energetic interactions between aminoacids, some of them would form strong and stabilizing bonds (in particular in the fold core), while other (on surface) would be involved in weaker contacts.

To characterize a protein fold, it is important to identify the crucial residues responsible for the energetic stability of the structure from those residues not forming stabilizing interactions.

To achive this goal, in our group has recently been proposed the energy decomposition method (EDM) [3,4,19,20] that, as a first step, computes the matrix of nonbonded interaction energies (namely, van der Waals and electrostatic interactions) between pairs of residues. This matrix is afterward diagonalized, and from the analysis of the eigenvector associated with the lowest eigenvalue, it is possible to identify those residues that behave as strongly interacting and stabilizing centers.

It is worthwhile to observe that there are two slightly different versions of the energy decomposition method. In the first one, the nonbonded interaction energy matrix is

obtained as an average over the MD trajectory. In addition solvation effects are not taken into account directly, although the solvent is present during the MD simulation and, hence, influences the protein structure.

In the second approach, after a cluster analysis performed on the MD trajectory, only the most representative structure of the most populated cluster is taken into account and the nonbonded interaction energy matrix is computed on that protein structure. Obviously, in the second case, the average solvent effect is not considered, and, to resolve this problem, the solvent is directly taken into account using the PBSA method [21,22] in the calculation of the nonbonded interactions.

It is important to observe that the two versions of the EDM provide qualitatively equivalent results, although the second one is less computationally demanding. Of course, because of the great number of simulations to be analyzed in our study, we opted for the second approach, using the GROMOS [15] methodology to carry out the cluster analysis with 0.2 nm as the root-mean-square deviation (RMSD) cutoff.

Now, for the sake of completeness, it is interesting to consider some theoretical details about the EDM. Let us indicate with M the non-bonded interaction energy matrix without the diagonal elements, namely without the self-interaction terms. This matrix can be diagonalized and expressed in terms of its eigenvalues and eigenvectors:

$$M_{ij} = \sum_{k=1}^{N} \lambda_k w_{ik} w_{jk} \tag{2.33}$$

where N is the number of aminoacids in the protein, λ_k is the k^{th} eigenvalue, and w_{ik} is the i^{th} component of the k^{th} eigenvector. Eigenvalues and eigenvectors are usually labeled following an increasing order. Therefore, λ_1 is the lowest eigenvalue, and hereafter, we will refer to the first eigenvector as the eigenvector corresponding to eigenvalue λ_1. The total nonbonded energy is defined as:

$$E_{nb} = \frac{1}{2} \sum_{i,j=1}^{N} M_{ij} = \frac{1}{2} \sum_{i,j=1}^{N} \sum_{k=1}^{N} \lambda_k w_{ik} w_{jk} = \frac{1}{2} \sum_{k=1}^{N} \lambda_k W_k \tag{2.34}$$

where $W_k = \sum_{i,j=1}^{N} w_{ik} w_{jk}$. If $|\lambda_1 W_1|$ is much larger than $|\lambda_k W_k|$ for $k \neq 1$, the sum over i and j of M_{ij} is dominated by the contribution due to the first eigenvalue and eigenvector, such that the total nonbonded energy can be approximated by

$$E_{nb} \approx E_{nb}^{app} = \frac{\lambda_1}{2} \sum_{i,j=1}^{N} w_{i1} w_{j1} = \frac{\lambda_1 W_1}{2} \tag{2.35}$$

As mentioned above, the eigenvector associated with the lowest eigenvalue is used to identify the most stabilizing aminoacids. In particular, considering its squared components as the weights of the corresponding residues in the structural stabilization, we can define "hot spots" as those residues with a weight higher than a threshold t. This threshold is set equal to the squared component of a normalized "flat eigenvector" (namely, a normalized vector whose components provide the same contribution for each site). This corresponds to a case in which each residue equally contributes to the structural stability, and therefore, the threshold t is equal to 1/N, where N is the number of the eigenvector components.

To sum up, with this analysis it is possible to identify aminoacids, involved in the most relevant non-bonded interactions, responsible for the stabilization of the protein fold.

Chapter 3

MLCE: a new method for epitope prediction

This chapter concentrates the development of a new method useful for the prediction of epitope locations on antigenic protein surfaces.

At the beginning of this chapter we explain the biological background concepts that are at the basis of this part of the thesis, in particular focusing on vaccine development and aspects related to protein antigens study. In the central part, the details of the methodology developed (called MLCE) and how energetic properties of a protein structure can be used to predict epitope residues are illustrated. A statistical analysis on nineteen different antigenic proteins has been carried out to test MLCE technique performance. At the end of the chapter, we describe BEPPE, a public web server implemented with MLCE method in order to make it available for the scientific community.

This chapter is based on the following scientific work:

Scarabelli G, Morra G, Colombo G. **Predicting interaction sites from the energetics of isolated proteins: a new approach to epitope mapping.** Biophys J. 2010 May 19;98(9):1966-75.

3.1 Biological framework

Development of antibiotic resistance in pathogenic bacteria is potentially one of the most serious threats in modern medicine [23]. The massive use of drugs and the possibility to exchange genetic material between infectious agents, resulting in acquisition of resistance and immunity, is causing a loss of effectiveness of commercially available antibiotics.

One approach to minimize the use of these medicines is to vaccinate population against strains of pathogenic bacteria, inducing and establishing protection against infective agents through the activation of the host immune system.

3.1.1 Immune response

The interaction between molecules is at the basis of cellular life. It is no surprise that the study of molecular recognition processes has emerged in recent years as a very important and relevant field of research, influencing areas like molecular biology, biochemistry, biomedicine, immunology, etc.

Regarding immunology, the recognition of antigens by antibodies is a fundamental step for the adaptative immune response which leads to the neutralization of a specific pathogen. Antigens are molecules (proteins or polysaccharides) expressed on the surface of infectious agents like viruses and bacteria which are recognised by the host immune system as non-self.

In case of infection, the immune system cells identify the invading agent through specific antigenic molecules expressed on it and activate the immune response. For the organism, this process is fundamental to avoid the spreading of the infectious pathogen and to neutralize it. The immune response consists at the beginning in the production of signalling molecules attracting cells like macrophages and granulocytes which start to fight the infection. Subsequently in the activation and differentiation of limphocytes B and T which can react specifically with the production of antibodies and the activation of cells able to kill the pathogen.

When the infection is defeated, immune cells (called memory cells) remain in circle in the host body to ensure a rapid and efficient immune response through the production of specific antibodies in case of a second infection from the same pathogen (Figure 3.1).

In this context, a vaccine is a biological preparation based on the capability of the immune system to keep memory of the infectious organisms. It is constituted by an inactivated pathogen, or a part of it, which is injected into an organism inducing protection against infective agents through the activation of the immune system and the production of memory cells.

In order to make a vaccine preparation it is important to know which pathogen molecules can be recognised by antibodies as not all the molecules expressed on a pathogen surface can be a good target for the antibody binding. In particular, the molecular regions recognized by antibodies assume huge relevance. Being able to characterize their chemical and physical properties and determine their presence and position on molecule surfaces, can result in a considerable advantage in vaccine design, reducing both industrial cost and production time [1].

Using high resolution structures, it will be possible to design antigenic molecules, making their production more efficient and improving the steps of the storing process, thereby lowering costs and eliminating problems to distribution. Furthermore, with the knowledge of which parts of an antigen must be retained to preserve basic char-

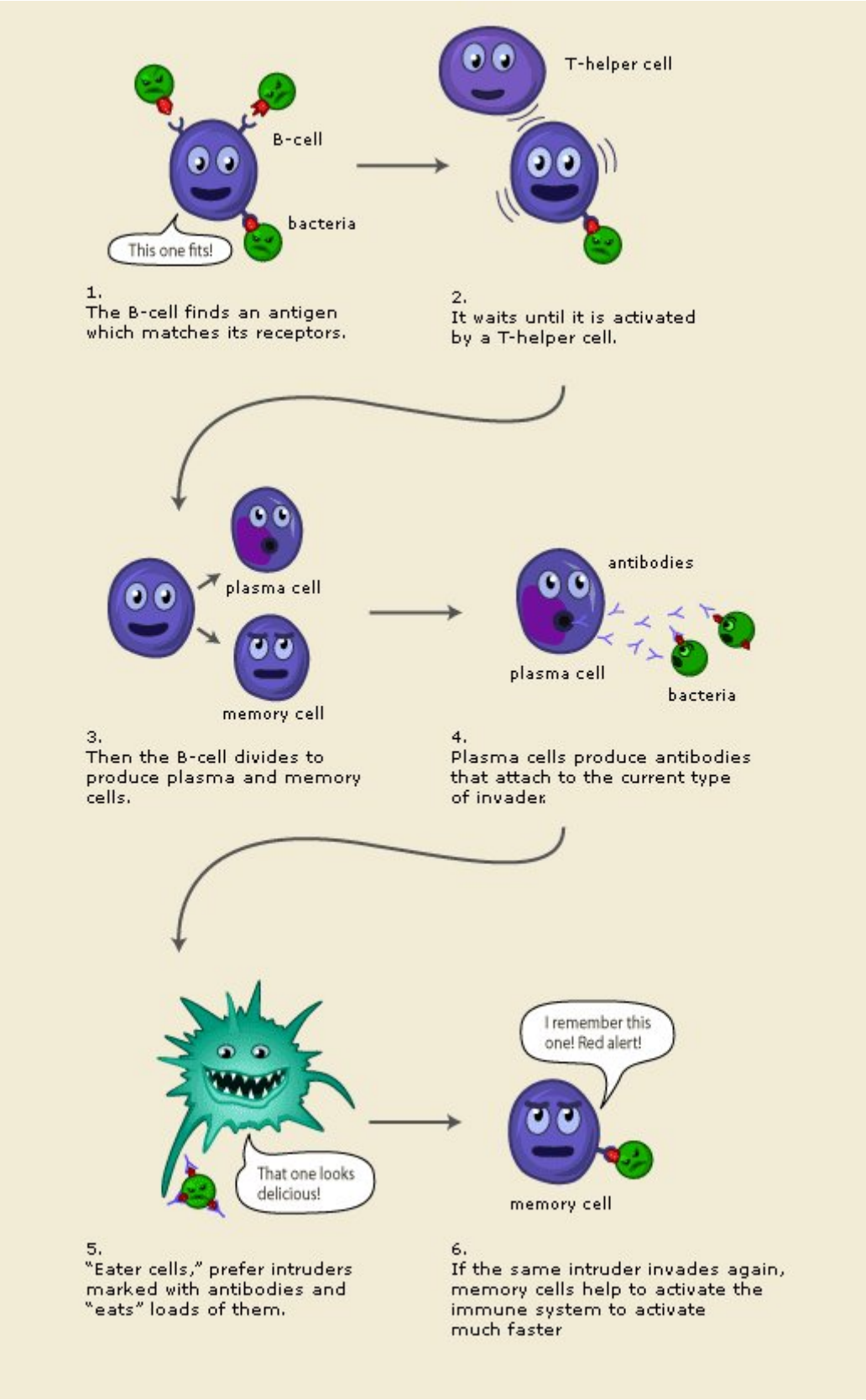


Figure 3.1: *Schematic representation of the humoral response. The antigens on the pathogen surface are recognized by specific B cell which start the production of antibodies in order to block the invading agent and to make possible its elimination. After defeating the infection, memory cells remain vigilant for further invasions from the same infective organism. Picture adapted from www.nobelprize.org*

acteristics for the recognition process and which can be altered, vaccine antigens can be modified more rapidly in response to changing epidemiology. Finally, the use of different molecules in combination vaccines and the immunization regimens can be simplified with a proper design.

All these considerations emphasize why it is very important to identify good antigenic molecules for the vaccine design. Antibody recognition is influenced by a lot of different aspects related to antigenic molecules like: size, shape, structural complexity, abundance, solubility, propensity to oligomerization, epitope structure and position and many others. Some of these aspects are in connection with the molecular structure of the antigen itself, in particular the ones related to the epitope regions. Indeed, the ability to predict epitope sequences and location is central to the development of structural vaccinology. Based on these aspects, the study of the structural and dynamical properties of these molecules, using computational techniques, can be a good instrument to improve the knowledge in this field.

3.1.2 History of vaccines

Vaccines were developed for the first time by Edward Jenner in 1796 starting from what a milkmaid boast said about human smallpox. She told Jenner she would never have the often-fatal or disfiguring disease smallpox, because she had already had cowpox, which has a very mild effect in humans. Jenner decided to verify this fact inoculating cowpox to a 8 years old boy. After 6 weeks he variolated the boy's arm with smallpox and observed that the boy did not catch smallpox.

In nineteen century Louis Pasteur generalized Jenner's idea by developing what he called a rabies vaccine (now termed an antitoxin) inactivating the pathogen organisms and inject them into patients.

Since those times, a lot of different vaccines have been developed using Pasteur methodology (or similar ones), which consisted in the use of killed or attenuated microrganisms, or with the selection of part or subunits of the agents like toxins or specific proteins. Examples of diseases overcome with these techniques are cholera, bubonic plague, hepatitis A, etc.

Unfortunately, for some pathogens all these kinds of approaches did not work, and so it was not possible to find a reliable vaccine to prevent their related diseases. This problems raised the necessity to find new techniques to identify good antigenic candidates to develop new vaccines.

3.1.3 Reverse Vaccinology

In the last decade the advent of post-genomic methodologies has had a great impact on immunology studies. The case of *Neisseria meningitidis* illustrates the strategic/synergic

role played by new genomics and bioinformatics methods in vaccine development and the possibilities they are bringing to overcome all difficulties, present in classical approaches, related to the design of vaccines.

Neisseria meningitidis is a major cause of bacterial septicemia and meningitis in humans [24], it is a Gram-negative bacterium, capsulated in its invasive form, classified into five major pathogenic serogroups on the basis of the chemical composition of distinctive capsular polysaccharides [25]. For one serotype (group B, called MenB) this polysaccharide is identical to a widely distributed human carbohydrate, making its use as the basis of a vaccine for prevention of MenB diseases problematic [26]. As consequence, most efforts have turned to development of vaccines based on surface-exposed or exported proteins [27]. Unfortunately, all the attempts tried in 40 years with classical vaccine development approaches failed, because of hypervariability and/or poor conservation among the diverse MenB strains that cause endemic disease, making it unlikely that they would provide broad protection against it [28]. For these reasons currently there is no commercial vaccine against serogroup-B *N. meningitidis*.

In 2000, when *N. meningitidis* genome was sequenced, a new technique to obtain new vaccine candidates emerged. The general idea is that a genome sequence containes the information of the complete repertoire of its antigenic proteins, so through a rapid and intelligent screening selection of all the organism proteins it can be possible to identify good antigenic proteins. This technique was named Reverse Vaccinology (RV, Figure 3.2) as it is based on a reverse approach compared to the classical vaccine development methodologies. Serogroup B of *Neisseria meningitidis* was the first case studied with the reverse vaccinology approach.

Starting from the sequenced genome of the MC58 MenB strain 2158 predicted ORFs (Open Reading Frame, one ORF corresponds to a region of the genome which is translated into a protein) were screened for the design of a novel vaccine [29]. Assuming that surface exposed antigens are the most suitable vaccine candidates, due to their potential to be recognized by the immune system, the draft MC58 genome was screened using bioinformatics tools, leading to the identification of 570 ORFs predicted to encode either surface exposed or secreted proteins [30, 31]. Antigen selection then continued based on the following criteria:

- the ability of candidate proteins to be cloned and expressed in *Escherichia coli* as recombinant (350 candidates)
- the confirmation of surface exposure through ELISA and FACS analysis
- the protective immunity obtained from the elicited antibodies, measured with serum bactericidal assay and/or passive protection in infant rats (28 candidates)

- the screening to determine the conservation level of the candidates within different meningococcal strains, primarily containing disease-associated MenB strains.

After all these steps few antigens have been identified by reverse vaccinology:

- genome-derived *Neisseria* antigen 1870 (GNA1870; which is factor H-binding protein [fHBP])
- GNA1994 (which is NadA)
- GNA2132 (GenBank accession number NP_275117)
- GNA1030 (GenBank accession number AAF41429)
- GNA2091 (GenBank accession number NP_275079)
- an outer membrane vesicles (OMV) from New Zealand MeNZB vaccine strain, which contains the immunogen PorA.

The vaccine formulation obtained is currently being tested in clinical phase III and consists of a combination of fHBP-GNA2091 fusion protein, a GNA2132-GNA1030 fusion protein, NadA and OMV.

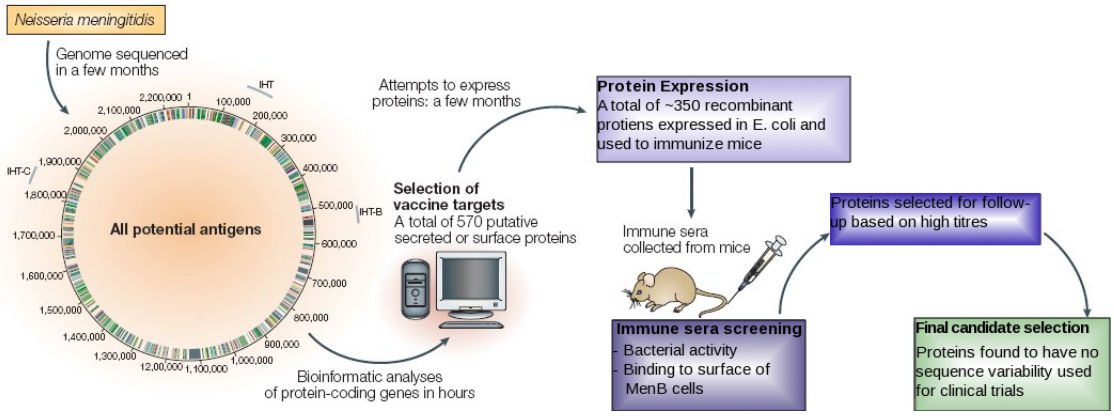


Figure 3.2: *This picture represents the reverse vaccinology main steps carried out on Neisseria meningitidis. Starting from the sequenced bacterial genome, more than 500 surface proteins have been selected, then ~ 350 have been expressed in Escherichia coli and after in vitro and in vivo assays only few of them were selected as vaccine candidates.*

The multivalent vaccine approach was taken due to the antigenic diversity of disease-causing MenB strains and should strengthen the protective activity of the vaccine, increasing in this way the breadth of MenB strains targeted by the vaccine and preventing the selection of escape mutants (i.e., bacteria that have a mutation in a gene encoding an antigen that would allow them to escape killing or neutralization

by vaccine-induced antibodies). When tested against a panel of 85 meningococcal isolates (predominantly MenB isolates) representative of the global population of disease-causing strains, the vaccine induced bactericidal antibodies in mice against 78% and 90% of strains when administered with the adjuvants aluminium hydroxide and MF59 (an oil-in-water emulsion), respectively [32]. Initial phase II clinical results in adults and infants indicated that this vaccine was well tolerated and induced a protective immune response against three diverse MenB strains in 89%-96% of subjects following three vaccinations and 93%-100% after four vaccinations [33].

This initial success of reverse vaccinology in developing a vaccine for MenB served as a proof-of-concept for this approach and catalyzed a paradigm shift in vaccine development. In fact this technique is independent of several of the constraints of classical vaccinology, such as the need to culture the pathogen in vitro (and all the difficulties related with). Today, the same genome-based approach is being applied to streptococci, chlamydia, staphylococci, *Plasmodium falciparum* and to bioterrorism associated agents such as *Yersinia pestis* [34]. These projects are based on stand-alone reverse vaccinology approaches, with some improvements like the use of pangenomics, and/or a combination of genomic, proteomic, and transcriptomic approaches.

3.1.4 Importance of epitope prediction

The success of reverse vaccinology has brought light on a fundamental question: why are some pathogenic proteins able to induce the production of bactericidal antibodies and many others not? Which are the necessary characteristics an antigenic protein should have to be recognised by an antibody?

Answer to these questions can be very useful to develop a rational vaccine design, improving the selection processes to obtain good vaccine candidates. Indeed, the optimization of vaccine antigens by combining the tools of genetic engineering with the insights provided by high resolution structural analysis will result in important advantages.

For example from the point of view of industrial applications, knowing the optimal structural characteristics for antibody binding can be useful to design and produce vaccines with higher efficiency and to improve the storage stability, resulting in the reduction of costs and in the elimination of problems related to the distribution. In addition, by engineering antigens amenable for use in combination vaccines, immunization regimens can be simplified. Also by knowing which parts of an antigen must be retained to preserve basic characteristics and which can be altered, vaccine antigens can be modified more rapidly in response to changing epidemiology [35].

All these aspects are related to the location of the antibody binding region on the antigen surfaces, so it is very clear why the possibility to predict the location of epitopes constitutes an important advantage to vaccine design.

In the effort to predict the antibody binding sites on antigenic proteins we have decided to study the structural properties of 19 antigens present in the Protein Data Bank (PDB). Importantly for all those proteins the structure of the isolated antigen and the structure of the antigen in complex with at least one specific antibody were present in the PDB.

Most of the computational methods available so far to predict the epitope positions was built only to locate linear epitopes (so epitopes constituted by residues consecutive in the protein sequence). Just few of them could identify structural regions, which constitute around 90% of all epitopes known [36]. Among this group many epitope predictors were built using information contained in data sets of known epitopes and no one considered the dynamical properties of the antigenic proteins. But as each molecule is subjected to movements and interactions with other molecules relevant knowledge can be lost without considering the dynamics.

For all these reasons we have thought that molecular dynamics simulations carried out on antigenic proteins could be very useful for the localization and characterization of epitopes as it takes into account the dynamical behaviour of a molecule. Several molecular dynamics simulations have been carried out on the isolated antigens. The knowledge of the epitope locations (obtained from the complex structures) has been used only to verify the predictions resulting from the MD calculations and not to make the epitope predictions themselves.

3.2 MLCE method

Epitopes are parts of the protein that can be recognized by a binding partner. Their sequences are typically mutation-tolerant [37], suggesting that they are not involved in the stabilization of the antigen fold. These sites have evolved, and must continuously evolve, to escape recognition by the host immune system, without impairing the native structure of the protein necessary for function in the pathogen [35]. Moreover, epitopes can be flexible and easily undergo conformational fluctuations [38, 39]. In other words, they are not involved in major intramolecular stabilizing interactions with other residues of the protein important to preserve the fold.

From the conformational and topological standpoints, epitopes are exposed regions on the protein surface, accessible for antibody binding [40]. Moreover, specifically in the case of discontinuous epitopes, high-resolution x-ray structures of antigen-antibody complexes showed they consist of residues whose spatial proximity relationships define a (large) patch on the surface of the antigen [41, 42].

Based on these considerations, we have set out to combine an analysis of protein

energetics obtainable from MD simulations with the topological information obtainable from the contact matrix of the representative structure of the trajectory [43]. The aim is to identify contiguous regions in the three-dimensional conformation of the antigen that are minimally coupled to the rest of the protein, and are thus likely sites for the dynamic modulation that would play a role in recognition events. The analysis of energetics is based on the energy decomposition method already developed in our group, which allows the detection of residue-residue couplings that are important in the stabilization of a fold (see details in previous chapter, energy decomposition method).

The method provides a simplified view of residue-residue pair interactions, extracting the major contributions to energetic stability of the native structure from the results of all-atom MD simulations. For a protein of N residues, the NxN matrix (M_{ij}) of average nonbonded interactions between pairs of residues can be built by averaging over the structures visited during an MD trajectory [2–4, 19, 43]. The rather noisy energy matrix is then simplified through eigenvalue decomposition.

Analysis of the N components of the eigenvector associated with the lowest eigenvalue was shown to identify residues that behave as strong interaction centers. These interaction centers are themselves characterized by components that have an intensity higher than the threshold value, which corresponds to a flat normalized vector with residues that would all provide the same contribution. It has been verified that applying this analysis to the representative conformation of the most populated structural cluster from the simulation yields the same results as the averaging over the equilibrated part of the trajectory [20].

As a caveat, it is worth noting that the latter approximation is valid when the most frequented cluster is significantly more populated than the others, so as not to neglect significant structural deviations captured by other clusters. In all the cases studied here this holds true, as we did not observe any major domain rearrangements, domain motions, or folding-unfolding events during simulations. The method was validated against experimental data and a relationship was found between the topological and energetic properties of a protein and its stability [2–4, 19, 43].

The map of pair energy-couplings filtered with topological information can be used to identify local couplings characterized by energetic interactions of minimal intensities. Because low-intensity couplings between distant residues in the structure are a trivial consequence of the distance dependence of energy functions, local low-energy couplings identify those sites in which interaction-networks are not energetically optimized. These regions may be regarded, therefore, as prone to interact with binding partners or to otherwise tolerate mutations that would preserve the antigen three-dimensional structure.

Moreover, thanks to the lower intensity constraints to the rest of the structure, these substructures would be characterized by dynamical properties that allow them to visit multiple conformations, a subset of which can be recognized by the antibody to form

a complex. The sites identified are typically clustered at the protein surface and are easily accessible. These concepts are somewhat reminiscent of local frustration, in which highly frustrated regions are often localized near interaction sites on protein surfaces [44].

3.2.1 Analysis of energetics and topological properties

The MLCE technique is based on the calculation of two matrices. The first is a filtered energetic matrix obtained with energy decomposition methodology previously explained (see section 2.2.3, structural properties). The second one is a contact matrix which takes into account information related to spatial proximity between each pair of residues.

The Energy Decomposition Method relies on the calculation of the interaction matrix M_{ij}, which is determined by evaluating average, interresidue, non-bonded (van der Waals and electrostatics) interaction energies between residue pairs, calculated over the structures visited during an MD trajectory. For a protein of N residues, this calculation yields an NxN matrix (where N is the residue number). As stated above, the same results can be obtained by calculating the interaction matrix M_{ij} from the representative conformation of the most populated cluster, in the absence of major conformational changes.

The aim of the energy decomposition method is to obtain a simplified picture of the most relevant residue-residue interactions in a certain fold and the interaction matrix can be represented as in equation 2.35.

From the physical point of view, this approximation indicates that any two residues i and j interact with energy $\lambda_1 w_i^1 w_j^1$. The value represents λ_1 a coupling parameter: a modulation of its intensity, as a result of mutations, can be interpreted as a rescaling in the intensity of protein interactions. A variation in the eigenvector components is related to a reorganization of native interactions that would modulate the contribution of a certain pair to the overall stability.

The principal eigenvector (defined as the sequence eigenvector, SE) constitutes a simple vectorial representation of the sequence: it reports on the contribution of each residue in the stabilization of the fold, which ultimately depends on the chemical properties of the residue itself. From this we can recover an approximation to the global stabilization energy, E_{nb}^{app} , which was shown to correlate with the relative different stabilities of mutants of several proteins, proving thereby to be a sufficient energetic descriptor to discriminate among them [43]. This method provides information on the mean coupling energy between two residues in the native state, revealing the network of most interacting residues through the structure.

The contact map of the representative structure from MD recapitulates which residue pairs are in contact in the conformation. If the distance between any two

C_{beta} atoms is below a cutoff value, the corresponding matrix entry is set to 1, otherwise it is set to 0. The distance cutoff is set to 6.5Å. For the sake of homogeneity with the energy matrix, contacts between nearest neighbors i, i+1 are included as well:

$$C_{ij} = \begin{cases} 1 & \text{if } r_{ij} < 6.5\text{Å} \\ 0 & \text{if } r_{ij} > 6.5\text{Å} \end{cases}$$

To calculate the contact matrices we consider the representative structure of the main cluster obtained with the GROMOS method from the MD trajectory of each antigen (cutoff value of 2Å [15]).

Energy decomposition was carried out both by averaging on structures saved every nanosecond during the simulations and on the representative protein conformation of the most populated structural cluster obtained from the trajectory. The resulting structures were minimized, and solvation effects were taken into account using the molecular-mechanics (MM)-Poisson-Boltzmann surface area (PBSA) method, with the nonbonded energy term for residues i and j [45] resulting in:

$$E_{ij}^{nb} = E_{ij}^{elect} + E_{ij}^{vdW} + G_{ij}^{solv} \tag{3.1}$$

3.2.2 Epitope Identification

The simplified interaction matrix defined by $\lambda_1 w_i^1 w_j^1$ is multiplied by the residue-contact matrix. This procedure allows to filter the information contained in the simplified energy matrix $\lambda_1 w_i^1 w_j^1$ in terms of residues that are close in space, highlighting pairs within the contact cutoff that are also coupled through nonbonded interactions. This provides a compact way to highlight which local pair-contacts in the three-dimensional organization of the protein are coupled through energetic interactions.

The resulting matrix can be viewed as the matrix of local coupling energies (MLCE, Figure 3.3). The contact-filtered coupling interactions are ranked in increasing order according to their respective intensities (from weaker to stronger). Starting from the minimum value (weakest local coupling interactions, defined as ”soft spots”, in contrast with the ”hot spots” characterized by high coupling intensities), the set of putative interaction sites was defined by including increasing residue-residue coupling values until the number of couplings that correspond to the lowest 15% of all contact-filtered pairs was reached. This then corresponds, in the approximation, to the set of local interactions, possessing minimal intensities, which may identify antigen-antibody or protein interaction sites. The corresponding residues define putative epitope sequences.

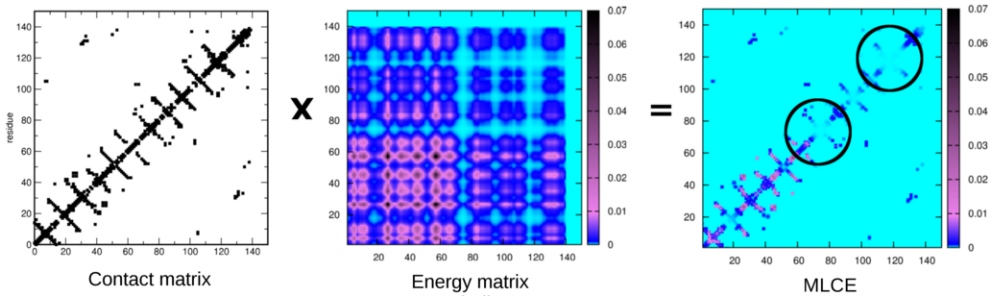


Figure 3.3: *Schematic representation of MLCE method. The contact map is multiplied by the simplified energy-coupling matrix. The resulting matrix reports the energetic coupling intensity of two residues in contact in space, represented as a color scale assigned to each point of the matrix. The weakest local interactions vanish in the background color: predicted epitopes are identified with circles.*

3.2.3 Simulation setup

All the starting structures of the proteins were downloaded from the Protein Data Bank, their respective codes are reported in Table 3.1 and subjected to explicit water MD simulations. The structures of the isolated antigens were solvated with explicit water, using the SPC water model [46] in triclinic or cubic boxes (depending on the protein shape) large enough to contain the whole protein and 1.4 nm of solvent from the protein.

Total charge of each system was neutralized with suitable counter ions. The systems were subsequently optimized through a molecular mechanics process with steepest descent algorithm (2000 steps). MD simulations were started from the minimized systems and carried out using GROMACS [47], with GROMOSS96 43A1 force field [48]. LINCS algorithm [49] was used to constrain the bond lengths for the all atoms. Electrostatic interactions were calculated through PME implementation of Ewald summation method, and temperature was set to 300K and kept constant with Berendsen thermostat [50] with a coupling constant of 0.1 ps. The timestep used was 2 fs.

Five simulations of 30 ns each with different random initial velocities were run for each protein, to check result dependence on simulation conditions. The first 5 ns of each simulation were discarded from final analysis. Average simulation time required using a parallel calculation on 32 Intel Xeon 3.166GHz cores for a protein of 200 residues has been around 24 hours.

3.3 Results

To evaluate the ability of our method to predict epitopes, we studied nineteen protein antigens for which crystal structures were available both in isolation and in complex with at least one specific antibody in the Protein DataBank (PDB). The dataset was

constructed by searching the PDB and initially discarding all complexes involving antibodies bound with only short peptide stretches, and focusing only on real protein-protein complexes.

Proteins studied

Antigen	Antigen-Antibody complexes	Biological role
1AO3	1FE8, 2ADF	von Willerbrand factor domain A3 (*H. sapiens*)
1AUQ	1OAK	von Willerbrand factor domain A1 (*H. sapiens*)
1BV1	1FSK	Major pollen allergen Bet V 1-A (*B. pendula*)
1CK4	1MHP	α I β I Integrin I-domain (*R. norvegicus*)
1CMW	1BGX	Taq DNA polymerase I (*T. acquaticus*)
1D7P	1IQD	Coagulation factor VIII precursor (*H. sapiens*)
1GWP	1AFV	Gag polyprotein (*HIV type I*)
1HCN	1QFW	Human chorionic gonadotropin (*H. sapiens*)
1K59	1H0D	Angiogenin (*H. sapiens*)
1KDC	1NSN, 2GSI	Staphilococcal nuclease (*S. aureus*)
1KZQ	1YNT	Major Surface Antigen SAG1 (*T. gondii*)
1P4P	1RJL	Outer surface protein B (*B. burgdorferi*)
1PKO	1PKQ	Myelin olygodendrocyte glycoprotein (*R. norvegicus*)
1POH	2JEL	Histidine containing phosphocarrier protein (*E. coli*)
1THF	1AHW	Human tissue factor (*H. sapiens*)
1UW3	1TPX	Prion protein (*O. aries*)
2VPF	1TZH, 1CZ8, 1BJ1, 2FJG, 2FJH	Vascular endothelial grow factor (*H. sapiens*)
3LZT	1IC4, 1NDG, 1DQJ, 1FDL	Lysozyme (*G. gallus*)
7NN9	1YQV, 1MCL, 1NDM, 1P2C, 2ZNW 1NCA	Neuraminidase N9 (*Influenza A virus*)

Table 3.1: *Table representing the selected antigens used to test MLCE method. All these proteins have different sequences, structures and are present in different organisms.*

Moreover, we selected antigen proteins whose structures had been solved in isolation via x-ray crystallography with a resolution lower than 2.6Å or via nuclear magnetic resonance (Table 3.1).

The set of isolated antigens was chosen to be nonredundant and diverse in terms of structures and sequences. Structural similarities were also minimal, because the antigens and epitopes were comprised from a diverse group of possible conformations that ranged from random loops to ordered secondary structures (Figure 3.4).

The predictive analysis was performed based only on MD simulations starting from the x-ray structure of the antigenic proteins in isolation. The validity of the epitope

prediction was benchmarked against the corresponding structure of the antibody-complexed antigen. MLCE method does not require the use of any training set of antibody-antigen complexes, as the determination of the epitope regions is based solely on structural-dynamical and energetic properties of uncomplexed antigens in isolation.

In Figure 3.4 it has been reported the projection of the low energy couplings on the surfaces of the proteins of the test set.

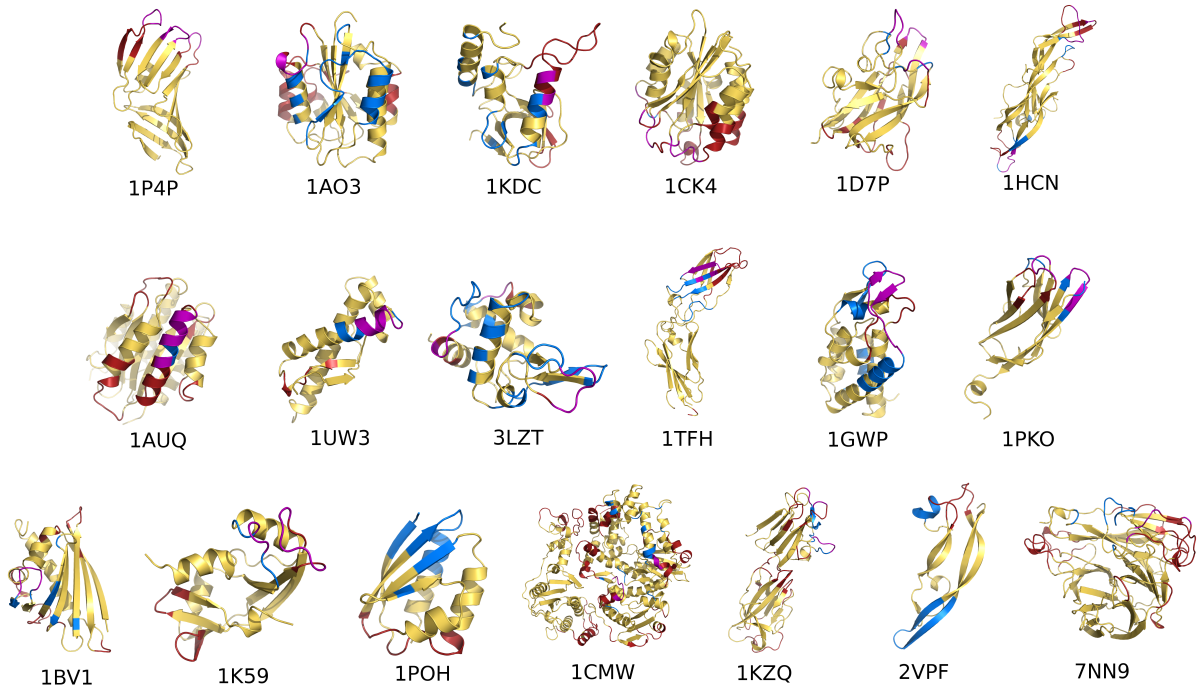


Figure 3.4: *Projection of the low-energy couplings from MLCE on their respective locations on the three-dimensional structure of all proteins analyzed. Predicted epitopes are in red, actual epitopes are in blue, and their intersection is in purple.*

3.3.1 Evaluation of epitope predictions

To assess the predictivity performance of the MLCE technique ROC curve statistical analyses on all antigens analyzed have been used. This analysis has been exploited previously in the immunoinformatics field in epitope prediction efforts [5, 38, 51] and is based on the calculation of four main parameters: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) [52]. The parameters are determined by comparing the predictions with experimental data.

For benchmarking, it was decided to rely on published articles on antigen-antibody complexes. We considered as an epitope the region beginning and ending with

aminoacids directly forming interactions with the antibody (defined by the crystal data). The epitope definition used includes also residues directly proximal in sequence with the previous ones, even though they may not directly contact the antibody in the complex x-ray structure, as they may have a relevant role in defining the optimal conformation required for recognition.

The parameters described above are used to determine different statistical measures:

$$FPR = \frac{FP}{TN + FP} \tag{3.2}$$

$$TPR \text{ (or Sensitivity)} = \frac{TP}{TP + FN} \tag{3.3}$$

$$Specificity = 1 - FPR \tag{3.4}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.5}$$

$$PPV = \frac{TP}{TP + FP} \tag{3.6}$$

The dependency of TPR versus FPR can be plotted in a graph known as the ROC curve. The area under the ROC curve (also known as the area under the curve, or AUC) is a good indicator of the performance of the method, and has been widely used in the evaluation of other approaches [38,53].

The points in these graphs have been determined by changing the cutoff threshold used on the low-energy contact matrix values to identify the possible epitope residues. To determine the ROC curve, we considered 19 different cutoff values on the ordered matrix elements, starting with the set of values containing the lowest 5% of the filtered contact energy values and increasing the threshold by 5% per step.

The area under the ROC curve, called AUC, is comprised between 0 and 1 (with a value of 0.5 for a random classifier) and it is useful to make comparisons among predictions obtained with different methods. AUC values have been calculated as the sum of the trapezoid areas determined by considering the points in the graph.

To assess the role of MD simulations, this analysis was carried out both on the representative structures obtained from the MD simulations and on single minimized

Performance of MLCE method in epitope predictions

Antigen	Antibody complexes	MD AUC	MM AUC	Sensitivity	Specificity	Accuracy	PPV	No. of epitopes	Interface residues
1AO3	1FE8, 2ADF	0.5	0.37	0.12	0.77	0.65	0.1	2	33
1AUQ	1OAK	0.89	0.95	0.78	0.79	0.79	0.15	1	9
1BV1	1FSK	0.64	0.41	0.35	0.78	0.74	0.16	1	17
1CK4	1MHP	0.88	0.71	0.92	0.78	0.79	0.22	1	12
1CMW	1BGX	0.64	0.62	0.30	0.77	0.76	0.05	1	30
1D7P	1IQD	0.82	0.85	0.67	0.80	0.79	0.30	1	18
1GWP	1AFV	0.58	0.55	0.30	0.90	0.72	0.58	3	47
1HCN	1QFW	0.81	0.73	0.54	0.86	0.81	0.38	2	28
1K59	1H0D	0.87	0.87	0.73	0.85	0.84	0.41	1	15
1KDC	1NSN, 2GSI	0.67	0.66	0.41	0.85	0.74	0.45	2	32
1KZQ	1YNT	0.56	0.61	0.36	0.74	0.70	0.12	1	22
1P4P	1RJL	0.98	0.98	1	0.89	0.90	0.48	1	13
1PKO	1PKQ	0.84	0.7	0.56	0.92	0.86	0.53	1	18
1POH	2JEL	0.25	0.24	0	0.83	0.62	0	1	21
1TFH	1AHW	0.69	0.66	0.30	0.82	0.75	0.21	1	27
1UW3	1TPX	0.94	0.91	0.67	0.95	0.92	0.62	1	12
2VPF	1TZH, 1CZ8	0.49	0.56	0.21	0.89	0.61	0.57	2	40
	1BJ1, 2FJG, 2FJH								
3LZT	1IC4, 1NDG, 1DQJ	0.67	0.66	0.2	0.92	0.56	0.72	3	65
	1FDL, 1YQV, 1MLC								
	1NDM, 1P2C, 2ZNW								
7NN9	1NCA, 1NMC	0.72	0.57	0.36	0.79	0.76	0.11	1	25
Mean		0.71	0.66	0.46	0.84	0.75	0.32	1.42	18.6

Table 3.2: *This table reports the results obtained with statistical analyses carried out on MLCE predictions. Columns 1 and 2: List of the Protein Data Bank (PDB) codes of the isolated proteins studied and used for prediction in this article and of the complexes with their respective antibodies, which were used for benchmarking. Columns 3-9: Area under the curve (AUC) values calculated with the matrix of local coupling energies (MLCE) approach on the structures obtained from extensive molecular dynamics (MD) simulations; from molecular mechanics (MM) minimization on the PDB structure; Sensitivity; Specificity; Accuracy; PPV; number of epitopes in the protein; number of residues in the epitopes.*

structures obtained directly from the PDB. Results are summarized in Table 3.2.

The MLCE approach provided good performances, with an average AUC of 0.71. In only one case the AUC value was lower than 0.5 (for Histidine-containing phosphocarrier protein, i.e., HPr, 1poh.pdb). In all other cases, MLCE determined putative antibody-binding sites with high ranking. One case of particular interest is lysozyme, where multiple antibody-binding regions are known (Figure 3.5a). MLCE approach proved able to identify these multiple binding sites. The same considerations can be applied to the cases of the von Willebrand factor A3-domain protein (vWF-A3; 1ao3.pdb; Figure 3.5b) and human chorionic gonadotropin (HCG, 1hcn.pdb; Figure 3.5c). Two

epitopes have been mapped on each protein and their location and sequences were correctly predicted by our approach.

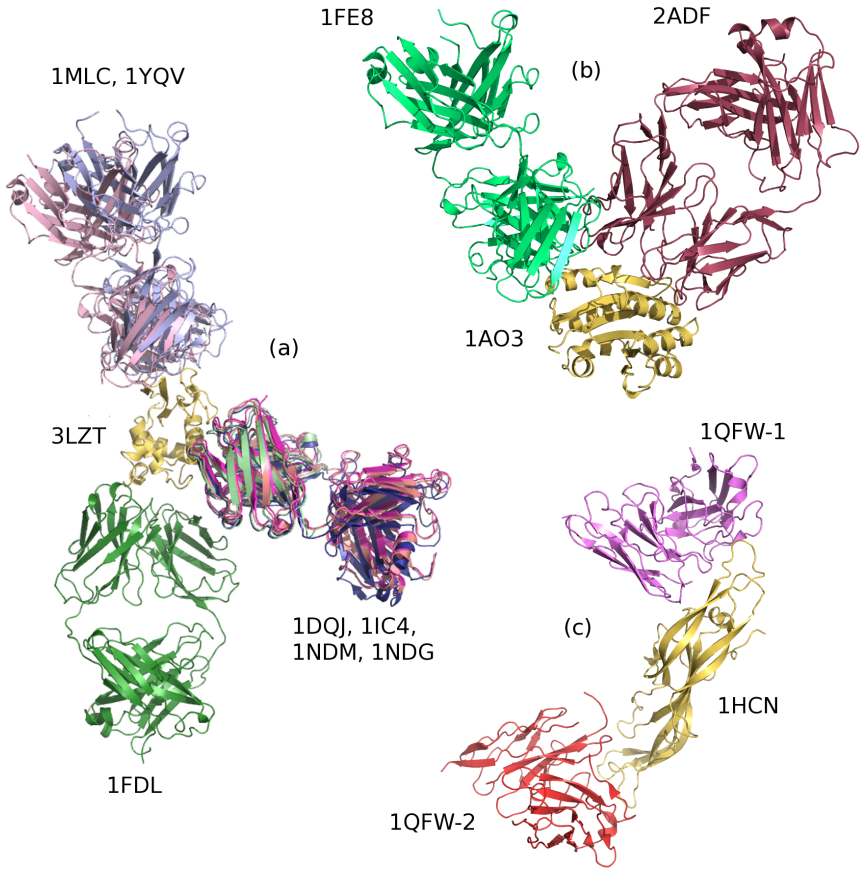


Figure 3.5: *Examples of multiple antibodies binding to the same antigen (light colored), highlighting the possibility for one protein to possess multiple epitopes. The PDB code of the antigen is close to the yellow antigen, and the PDB codes of the complexes are near each respective antibody. (a) lysozyme; (b) the von Willebrand A3 factor; and (c) human chorionic gonadotropin.*

In addition sensitivity, specificity, and accuracy of MLCE method have been evaluated, based on the definitions of Ponomarenko and Bourne [5]. Within the chosen threshold, the sensitivity (the proportion of correctly predicted epitope residues with respect to the total number of epitope residues) gave an average value of 0.46, which is slightly better or comparable to the values reported in the literature [38, 53]). The results of specificity (the proportion of correctly predicted nonepitope residues with respect to the total number of nonepitope residues) and accuracy (the proportion of correctly predicted epitope and nonepitope residues with respect to all residues) provided average values of 0.84 and 0.75, respectively (Table 3.2).

This confirms the applicability of the approach to the identification of putative binding-sites on protein surfaces.

Finally, the positive predictive value (PPV) of the method was determined. This value reports on the proportion of correctly predicted epitope residues with respect to

the total number of predicted epitope residues. The results obtained with MLCE are in line with the performances of several known predictors of protein-protein interaction sites and protein-protein docking programs reported in Ponomarenko and Bourne [38] and in de Vries and Bonvin [53].

Based on the comparison with other algorithms, at least one-half of the predictions may be useful to direct protein-protein docking efforts by reliably focusing on the predicted epitope region. Considering that antigens are notoriously hard to predict, this can be considered a positive result of the MLCE approach, given that it relies only on a general physical hypothesis for protein-protein interactions and on no previous assumptions regarding epitope sequences, shapes, etc.

3.3.2 Structural properties of predicted epitopes

The structural properties of predicted binding-sites were examined to evaluate the ability of the method to retrieve epitopes from any secondary structure motif, and to discriminate the antibody-binding properties of loops within the same structure.

The case of the OspB C-terminal fragment from Borrelia burgdorferi (1p4p.pdb) is particularly interesting (Figure 3.4). The antibody-binding region is defined by a discontinuous (conformational) epitope that consists of residues that belong to three different loops. The protein also presents several other loops. The MLCE method is able to discriminate the three loops making up the epitope region from the other loops. The epitope-loops are actually decoupled, in terms of stabilizing interactions, from the rest of the protein. The remaining loops provide stabilization energy to the folding core, and thus may not undergo conformational changes, interact with other proteins, or tolerate mutations without major energetic costs.

Importantly, MLCE could also detect epitopes that are part of ordered secondary structures. Epitopes with α helical structures are predicted for 1auq, 1uw3, and 3lzt. Epitopes in β sheet conformations are correctly detected in 1tfh, 1gwp, and 1pko.

In the case of human angiogenin (ANG, 1k59.pdb), MLCE identifies an additional region of low-energy coupling located at the opposite face of the molecule from the antibody-binding site. Indeed, crystal structure determination has shown that binding of the complementarity determining regions of the antibody induces a dramatic conformational change precisely at the region of angiogenin opposite to the epitope [54], which is used by the protein to bind to cells. Moreover, in the case of Histidine-containing phosphocarrier protein (HPr, 1poh.pdb), where the calculated AUC value from our calculation is as low as 0.25, the protein-region of lowest energy couplings coincides with the substrate-binding site.

As epitopes identified with MLCE are minimally coupled to the rest of the protein, they are also endowed with higher flexibility, as shown by the root-mean-square fluctuation graphs. Importantly, the analysis of flexibility profiles alone is not sufficient to

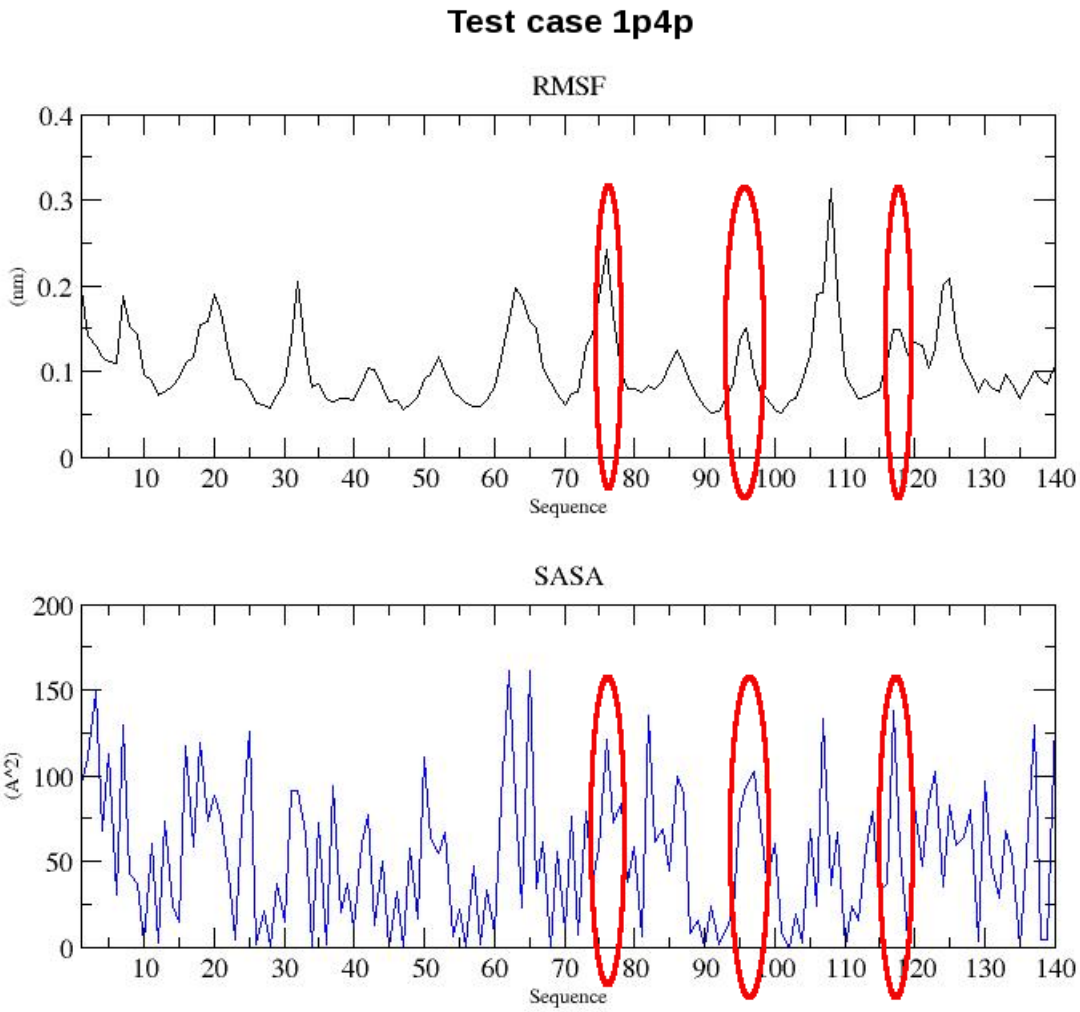


Figure 3.6: *Examples of fluctuations and SASA determined on a test case antigen (1p4p.pdb). Residues in the circled regions constitute the epitope, as it is shown it is difficult to set up a threshold to distinguish epitopes from other protein regions.*

discriminate between epitope and nonepitope regions (Figure 3.6, upper).

An attempt to characterize epitopes was carried out calculating the Solvent Accessible Surface Area (SASA) on the main representative structure of the protein using Naccess program [55]. SASA is determined moving a sphere with a radius of 1.4Å on protein structure surface profile. Unfortunately, this approach was not successfull to distinguish epitope sites from others like exposed loop regions (Figure 3.6,down).

Similar results have been obtained with Normal Modes calculation. In Figure 3.7 are reported, as an example, fluctuations of protein residues along the first three lowest energy normal modes. Normal modes have been obtained both by using an Elastic Network Model to represent protein structures [56] and by an all-atom representation of the protein implemented in Amber program. As shown in the graph, epitope identification is difficult as these residues behave in the same way as other protein

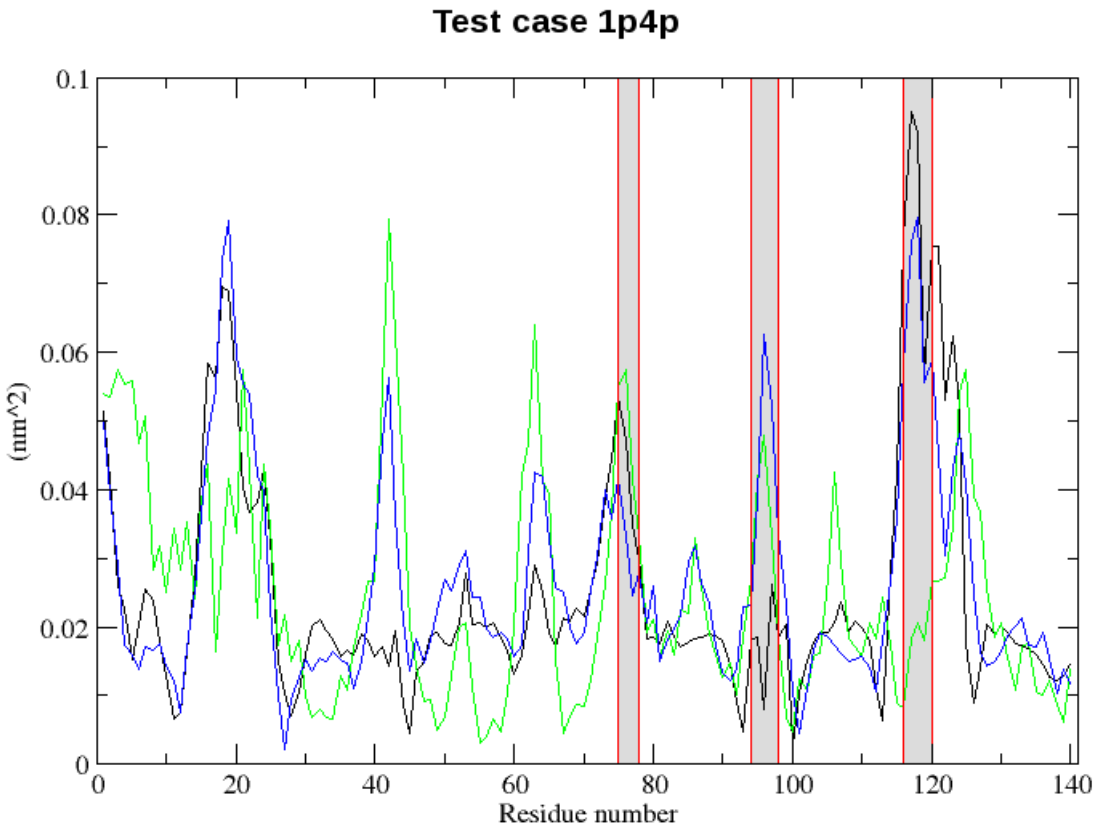


Figure 3.7: *Normal Mode Analysis carried out on 1p4p test case. In the graph are represented the first 3 normal modes calculated on the principal cluster structure. In gray shadows are highlighted epitope regions.*

regions.

These observations corroborate with the hypothesis that MLCE has the ability to detect sites poised to interact with other partners.

3.3.3 Impact of MD simulations on predictions

The use of MD simulations improved the results of our functional predictions. Indeed, the performance of the method appears to deteriorate slightly when applied to the structures of antigens extracted directly from the PDB, yielding an average AUC of 0.66 (Table 3.2).

Finally, we also tested the dependency of MLCE performance on the simulation length. To this end, each trajectory was split into 2ns intervals and the performance was evaluated on increasing time windows (Figure 3.8). In general, the performance, in terms of the resulting AUC value, converges within the first 4-6ns showing a possibility that one might employ shorter simulation times than those proposed here. In any case, as the final goal is to make available the method in a server-based version useful predictions can also be obtained with the use of the simple MM-PBSA approach.

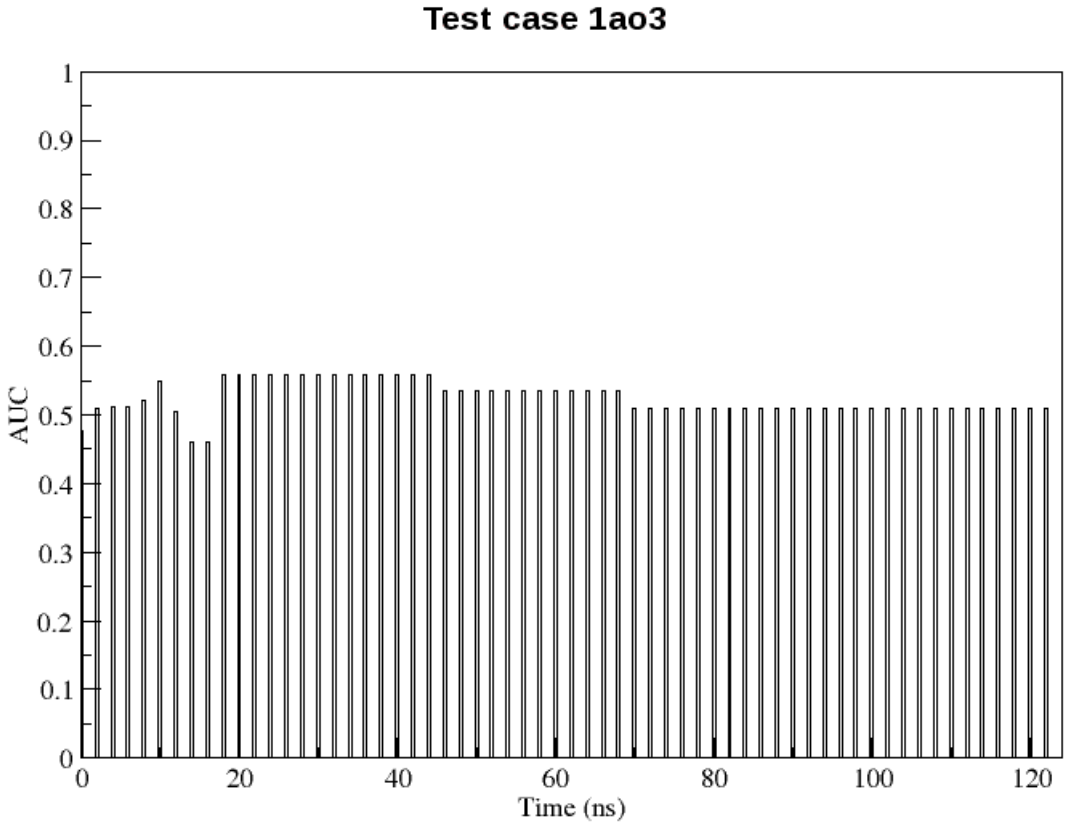


Figure 3.8: *Example of AUC values obtained cutting recursively MD trajectory every 2ns. MLCE calculation has been carried out on the truncated simulations. Cut-time is represented on the x-axis, while AUCs on the y-axis.*

3.4 BEPPE (Binding Epitope Prediction from Protein Energetics), a new web server for epitope prediction

A public web server, called BEPPE (available at, http://158.109.215.216/upload.php?UserName=103) was subsequently implemented with MLCE method. The aim is to carry out epitope predictions on single protein structures obtained from X-Ray, NMR or through computational tools like homology modelling or MD simulations. Screenshot of the main page is represented in Figure 3.9.

To use BEPPE the input steps required are: upload a protein structure in PDB format, select the softness level of the prediction and insert an email address to receive the output.

3.4.1 Input

BEPPE is implemented to allow the user to choose the prediction softness, meaning that it is possible to consider for epitope identification the 10% (strict), 15% (defalut) or 20% (soft) lowest energetic residue couplings in MLCE calculation. As result number

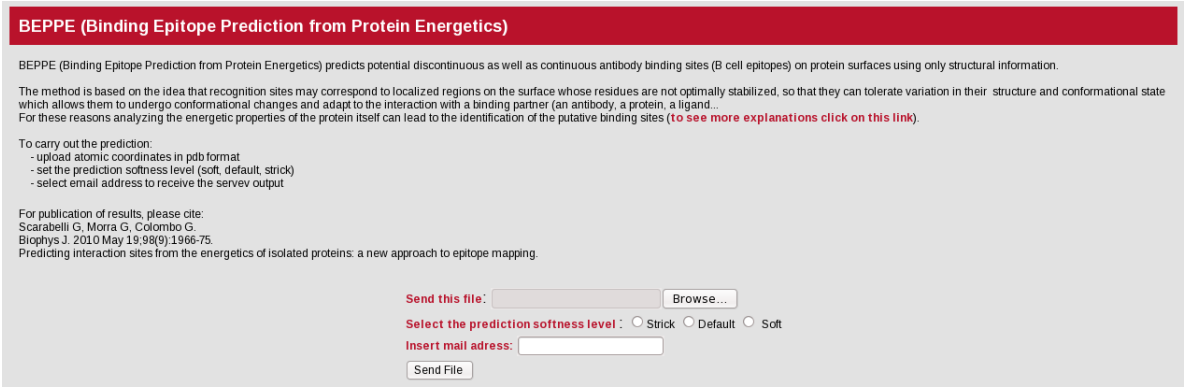


Figure 3.9: *BEPPE main page screenshot. A brief explanation is on the page top (with a link to another page containing more details), then the instructions to run a calculation are reported and at the bottom the reference scientific article is written.*

of predicted residues and patch sizes will vary according to user selection (going from few residues with a strict prediction to a higher number with a soft one).

3.4.2 Method

BEPPE reads the input and then performs the following steps:

- Optimize protein structure through 500 steps of molecular mechanics
- Perform MLCE calculation
- Select predicted residues and clusterize them into patches using the contact matrix previously calculated, rank the patches according to their average energies and select up to 4 different patches (discarding the ones composed by less than 5 residues)
- Create sequence segments in fasta format (1 letter code) joining predicted residues which are far away up to 4 positions in sequence
- Align these segments (using BLASTP algorithm [6]) with all human protein sequences available in order to find mimotopes (a mimotope is an epitope which is homologue to a region present in a host protein, making it problematic to be recognised by host antibodies [7])
- Send the output to the email address specified by the user

3.4.3 Output

The output consists in:

- predicted residues clusterized into patches
- a Pymol script readable with Pymol program, useful to have a three-dimensional representation of the protein with patches highlighted
- results of the alignment with human proteins
- a link to download output files

BEPPE performance was compared with other three different epitope predictor web servers, DiscoTope [57], Bepro [58] and Ellipro [5].

DiscoTope predictions are based on aminoacid statistics, spatial information, and surface accessibility in a compiled data set of discontinuous epitopes determined by X-ray crystallography of antibody-antigen protein complexes. Bepro relies on aminoacid propensity scale along with side chain orientation and solvent accessibility information using half sphere exposure values. Both DiscoTope and Bepro are built using a training set of antigenic protein with known epitopes. On the other hand Ellipro is not knowledge based, it uses geometric information contained in the structure, approximating protein shape as an ellipsoid and determining a protrusion index for each residue. Resulting aminoacids are then clusterized based on their protrusion index values.

Server performances on the nineteen antigen proteins used to test BEPPE are shown in Figure 3.10. PPV and sensitivity have been selected to evaluate the different approaches. As shown in the figure, BEPPE presents a slighly lower value for sensitivity while it has a better value for PPV compared to the other methods. The clusterization step, following MLCE calculation, reduces the number of predicted residues, improving the PPV average value shown in Table 3.2 but decreasing the average Sensitivity value.

In conclusion, BEPPE has the advantage of being able to find structural epitopes without using any antigenic training set, it is not knowlegde based and epitopes are predicted just relying on physical and chemical properties of the protein itself. In addition the quite remarkable PPV average value (0.44) underlies that almost the half epitope residues predicted by BEPPE are part of real epitopes.

3.5 Discussion

Reliable prediction of antibody-binding sites for a specific protein is a necessary condition to the discovery of new therapeutic opportunities in immunology. One fundamental aim of structural vaccinology is the selection of protein candidates with optimized properties in terms of sequence, structure, and presentation of the determinants for antibody-recognition.

In this context, upon being conducted on new pathogens, high-throughput genomic investigations (such as those employing RV) may reveal target antigens that have little

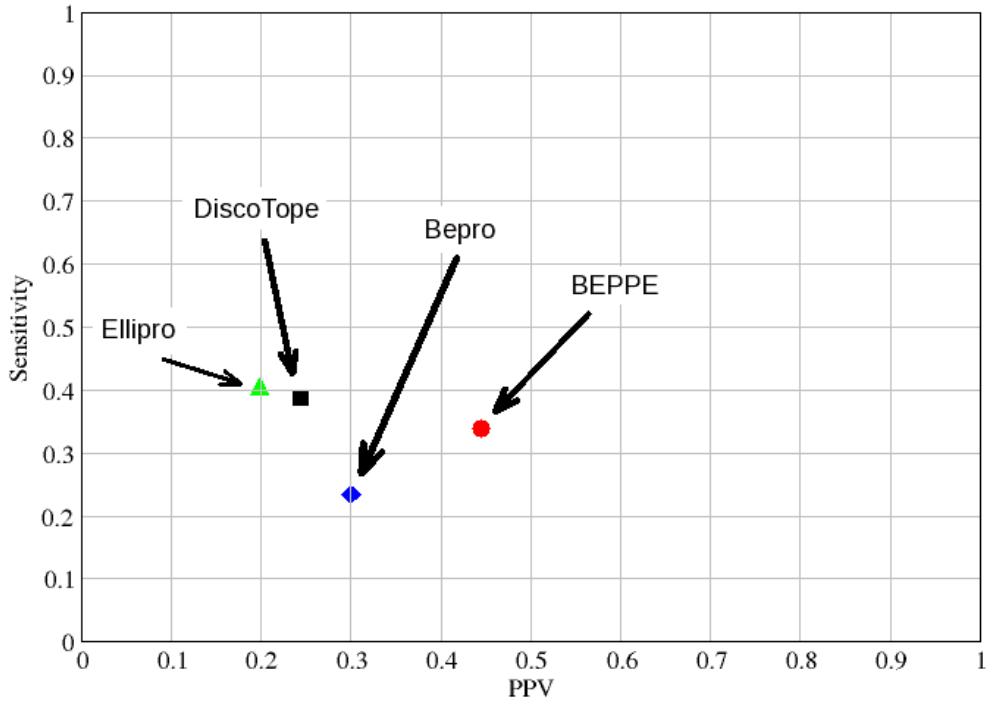


Figure 3.10: *Sensitivity and PPV calculated on the principal cluster structures of the nineteen antigens reported in Table 3.2 have been used to compare different techniques to predict epitopes. As shown BEPPE reaches almost the same values for sensitivity but higher values for PPV.*

sequence similarity to functionally annotated ones, and which may contain novel folds. Consequently, it is important to develop computational methods that can help identify potential epitope regions of an antigen independently of its sequence and/or shape similarity with other known proteins, and independently of the knowledge of related structures of antibody complexes.

Starting from these considerations, in this thesis work, the development of a new approach for the identification of possible antibody binding sites based uniquely on the structure, dynamics, and energetics of the protein-antigen in isolation is reported. The corresponding antibody-bound complexes are used for benchmarking the results.

The approach proposed is based on simple energetic and conformational concepts. Antigenic proteins must fold to a well-defined three-dimensional structure to properly carry out their functions in the pathogen.

The stabilization of the folded state can be achieved through interactions of higher intensity between specific residues that define the folding nucleus. Mutations in the folding nucleus have been shown to impact on protein stability and foldability [2,4,19,

20, 59–61]. In contrast, epitopes are typically mutation-prone sites [35, 62]: a protein from a pathogen should, in fact, be able to tolerate mutations that could help it evade the immune defense system of a host.

The energy decomposition method that has been introduced and tested proved able to single out the residues of the folding nucleus and flag their contribution to stabilization energy [2,20]. The ability to identify the folding nucleus complementarily determines the possibility to identify positions that are more tolerant to mutations. Typically, they coincide with the residues characterized by low energetic couplings with the remainder of the protein. Moreover, low-intensity couplings between proximal residues define sites whose interaction-networks are not energetically optimized and which are generally located on the surface.

From the dynamic point of view, these substructures may easily undergo conformational transitions and fluctuations favoring the docking of potential binding partners through a conformational selection mechanism [63]. Binding of a specific antibody partner would thus select specific geometries of the antigen, shifting the equilibrium toward thermodynamically stable complexes.

Based on these premises, the positioning of these sites can be identified in a compact way by multiplying the simplified energy-coupling matrix by the residue-contact matrix. This procedure allows to filter the information contained in the simplified energy matrix in terms of residues that are close in space, highlighting pairs within the contact cutoff that are also energy-coupled through nonbonded interactions in the three-dimensional structure. By concentrating on the lowest energy-coupled pairs in contact according to the contact matrix definition, it is possible to identify surface patches that can be recognized by a putative binding partner.

It is important to underline that the aim is to identify, specifically, locally organized residues with nonoptimized interaction energy-networks that are independent of possible dynamic signatures. Interestingly, analysis of normal modes or cross-correlation coefficients of residue pairs could not identify any specific, nonrandom fluctuations involving spatially localized regions with the lowest energetic-couplings.

Epitopes are characterized by (anti)correlated as well as random motions with the rest of the protein or with other parts of the conformational epitope. This aspect can be interpreted in the light of the weak energy-couplings among epitope residues, which result in higher flexibility and in the absence of major conformational constraints to the rest of the protein.

The surface patches identified through our procedure define the three-dimensional, structured landscapes associated with discontinuous epitopes that are recognized by antibodies.

It is worth noting, once more, that the whole procedure for epitope identification is based on the study of the antigen in isolation, and the structures of antigen-antibody complexes are used only as a posteriori validations of the analysis. The predictivity,

specificity, and accuracy of the method are in line with what has been reported recently in the literature [38,53].

Interestingly, MLCE proves able to identify multiple epitope sites encompassing different antigen regions (Figure 3.5). A certain protein surface may in fact contain several possible antibody-binding sites that may not be represented in the sets of structures currently available. Spatially localized sites with low energetic coupling to the rest of the protein may determine the dynamics required for specific function and/or recognition of partners other than just antibodies [64].

In the case of angiogenin (ANG,1k59.pdb), in addition to correctly predicting the epitope, MLCE method detects the cell-binding region of angiogenin at the opposite part of the molecule from the combining site [54]. In the case of Histidine-containing phosphocarrier protein (HPr, 1poh.pdb), the regions of low-coupling energy include the phosphate-binding site, located in the N-terminal region.

Nonoptimized interaction networks can be exploited by the protein to modulate structural plasticity and local flexibility and provide conformational and functional adaptability to possible binding partners. As is also suggested by Ferreiro et al. [44], localizing alternate conformational states or sequence mutations on specific substructures, while minimizing the influence on the three-dimensional stability required for function, could provide a mechanism of specific control of motions by concentrating only on a subregion of the protein.

The MD-based method we propose is clearly not as efficient, in terms of computational expenses, as are dedicated bioinformatics tools and servers that build on different ideas [5,51,58,65].

In most cases, the use of MD structures (either the most representative conformation from cluster analysis or averaging over the trajectories) gives only slight improvements over direct minimization of the PDB structure of the antigen. MD-related performance improvements in our data set are noticed mainly for cases in which an epitope is shared between a secondary structure element and a loop.

The release of strain determined by the initial crystal packing and consequent conformational relaxation determined by MD favor the geometric and energetic organization optimized for the recognition of the binding partner. In this framework, the role of MD can be most relevant in cases where major structural rearrangements are involved, e.g., in domain motions, large conformational changes, and local folding-unfolding. These cases were not present in our initial dataset, but correct epitope predictions have already been obtained for multidomain aminoacid transporting proteins from the pathogen Chlamydia.

Given all these caveats, it is, however, important to underline that the aim of the study was to introduce a conceptually different approach. We notice that dramatic improvements in algorithms and hardware solutions [66–68] might make it possible to obtain large-scale MD-based predictions more quickly and on longer timescales.

Despite its limitations, we think that the MLCE method may be a valid tool to direct epitope-mapping experiments and possibly identify binding patches to restrict the search of binding poses in protein-protein docking algorithms. With regard to epitope mapping, our approach is already being applied to targets of industrial interest (see next section).

Further improvement of the predictions may be obtained by integrating MLCE with other predictors that are based on bioinformatics analysis.

From the point of view of possible applications, this method may be relevant for structure-based vaccine design. It is possible in fact to focus antigen mutagenesis on those regions that are not part of the folding core and, by so doing, preserve the fold and leave the three-dimensional structure of the protein and epitope presentation unchanged. Random or site-directed mutagenesis could thus be concentrated upon the putative epitope sites, eventually selecting new sequences with maximum affinity for neutralizing antibodies.

An alternative strategy would imply the stabilization of the structure of the antigen by engineering Cys cross-linking mutations, or by further optimization of the folding core, to obtain a dominant conformation that would stably present the antibody recognition determinants rather than transiently populate binding conformations.

Finally, by knowing which parts of the antigens can be modified and which should be left unchanged to retain efficient neutralizing antibody recognition, protein antigens could be modified and selected to optimize production and storage, with an impact on costs and distributions of potential vaccines.

Chapter 4

The Chlamydia ArtJ paradigm, an industrial test case

This chapter focuses on the use of structural data obtained on two antigenic proteins expressed on *Chlamydiae trachomatis* and *Chlamydiae pneumoniae*, called CT381 and CPn0482 respectively, for the identification and characterization of functionally immunogenic domains.

CT381 and CPn0482 are both present on bacterium surface and show a good homology between their sequences. Trying to characterize and compare their antigenic properties can result in an important improvement in vaccine design.

In the first part of this chapter a brief introduction explains the biological framework related to CT381 and CPn0482 antigenic proteins. Subsequently, we describe experimental and computational analyses carried out on them. In the end, data obtained with these different approaches are compared to completely characterize the protein targets.

This work has been carried out in collaboration with other research groups: Novartis Vaccine (Siena, Italy) has accomplished molecular biology and immunology analyses, BioXtal (Marseille, France) has resolved the crystal structure of the two proteins, Universitàt Autonoma de Barcelona (Bellaterra, Spain) has simulated dynamical behaviour of these antigens and Jacob University of Bremen (Bremen, Germany) and our group have predicted epitope residues for those proteins.

This chapter is based on:

Soriani M, Petit P, Grifantini R, Petracca R, Gancitano G, Frigimelica E, Nardelli F, Garcia C, Spinelli S, **Scarabelli G**, Fiorucci S, Affentranger R, Ferrer-Navarro M, Zacharias M, Colombo G, Vuillard L, Daura X, Grandi G. **Exploiting antigenic diversity for vaccine design: the Chlamydia ArtJ paradigm.** J Biol Chem. 2010 Sep 24;285(39):30126-38. Epub 2010 Jun 30.

4.1 Introduction

One of the riddles in vaccine research is that subunit antigens showing high sequence homology among bacterial species may eventually display diverse antigenic and protective properties.

For example, antigens belonging to very conserved protein families, such as those including glycolytic enzymes or heat shock proteins, are highly expressed and immunogenic in most bacterial pathogens but are found to be protective only for some of them [69]. Therefore, it emerges that sequence conservation and structural similarities are necessary but not sufficient prerequisites to predict antigen immunologic properties.

Experimental evidence that highly homologous proteins fail to elicit cross-protection against closely related heterologous strains may be ascribed to a variety of causes, including differential expression level and/or cellular localization. The in vivo capacity of an antigen to raise antibodies able to protect from bacterial infection, by either neutralizing bacterial entry or promoting their killing, could also depend on a complex combination of properties including structural complexity, dynamics and epitope distribution.

In this context, *Chlamydiae trachomatis* (CT) and *Chlamydiae pneumoniae* (CPn) are an optimal test case to study, as they show two homologue proteins (called ArtJ), expressed on their surfaces that, although similar at sequence level, show diverse immunogenic properties [70].

For this reason we performed structural and functional analyses of ArtJ orthologues to characterize their antigenic properties, trying to improve the research of a vaccine able to induce immunity against these pathogens.

C. trachomatis is an infective bacterium causing prostatitis and epididymitis in men while in women is responsible for cervicitis, pelvic inflammatory disease (PID), ectopic pregnancy and it is one of the most common sexually transmitted infections worldwide. On the other hand *C. pneumoniae* is a major cause of pneumonia, pharyngitis, bronchitis and atypical pneumonia.

ArtJ protein is so annotated by analogy with the ART transport systems of *E. coli*, which has five genes organized in two operons [71]: artPIQM and ArtJ, which are responsible for the arginine transport. In CPn, however, the artPIQM genes are absent and, therefore, it appears that chlamydial ArtJ operates in a molecular context that is different from the *E. coli* model and must be peculiar to this species.

Moreover, ArtJ is able to induce high antibody titers both in mouse models and human patients that experienced a *C. trachomatis* infection. However, while recombinant CPn ArtJ elicited antibodies able to neutralize Chlamydia infectivity the CT protein did not show this functional activity [70].

This evidence raised the question whether differences in structural features and

related properties such as dynamics, specific intramolecular interactions and electrostatics, between CT and CPn ArtJ may account, in addition to (or as a consequence of) sequence differences, for their different immunogenicity.

In this chapter the investigation of the antigenic properties of ArtJ in the two Chlamydia species will be illustrated by exploiting new structural information.

4.2 Experimental and computational procedures

CT and CPn ArtJ ORFs (Open Reading Frame) were PCR-amplified using their respective chromosomal DNA as template and amplification products have been used to transform cells in order to obtain their expression. The proteic extract was purified and then proteins was crystallized in order to determine their molecular folds. The structures have been deposited in Protein Data Bank (PDB) with the following entry codes: 3G41 (CPn ArtJ) and 3DEL (CT ArtJ).

In addition mouse antisera and *in vitro* assays have been performed in order to test production and binding ability of specific antibodies against ArtJ proteins.

For further details on these parts see Soriani M. et al. J Biol Chem. 2010 Sep 24;285(39):30126-38.

Molecular dynamics simulations were carried out on the crystal structures obtained. All MD simulations were performed with software package GROMACS 3.3.1 using GROMOS 45a3 force field with SPC water model, and periodic boundary conditions. Temperature and pressure were kept respectively at 300 K and 1 bar.

Simulations of CT ArtJ and CPn ArtJ included residues from GLU34 to ASN257 and from ARG31 to GLU255, respectively. The initial open-apo conformation of CPn ArtJ, and the initial closed forms (both apo and bound) of CT ArtJ, were obtained by homology modeling using the MODELLER [72] program with CT ArtJ and CPn ArtJ crystallographic structures used as template. This procedure generated six different systems for simulation.

Protein structures taken every 25 ps in the period 100-200 ns of the five simulations performed for each system were clustered with the algorithm proposed by Daura [15], using a cut-off of 0.2 nm and all C-α atoms (including that of the arginine ligand, when present) for the calculation of RMSD. Then the resulting largest structural cluster obtained in each case was used for energy-based predictions of epitope locations.

4.3 Structure of ArtJ proteins

To investigate the presence of structure-dependent factors that could be partly accounting for the different antigenic properties of the two orthologs, the crystallographic

structures of CT and CPn ArtJ were determined by X-ray crystallography at 1.9Å and 2.1Å, respectively (Figure 4.1).

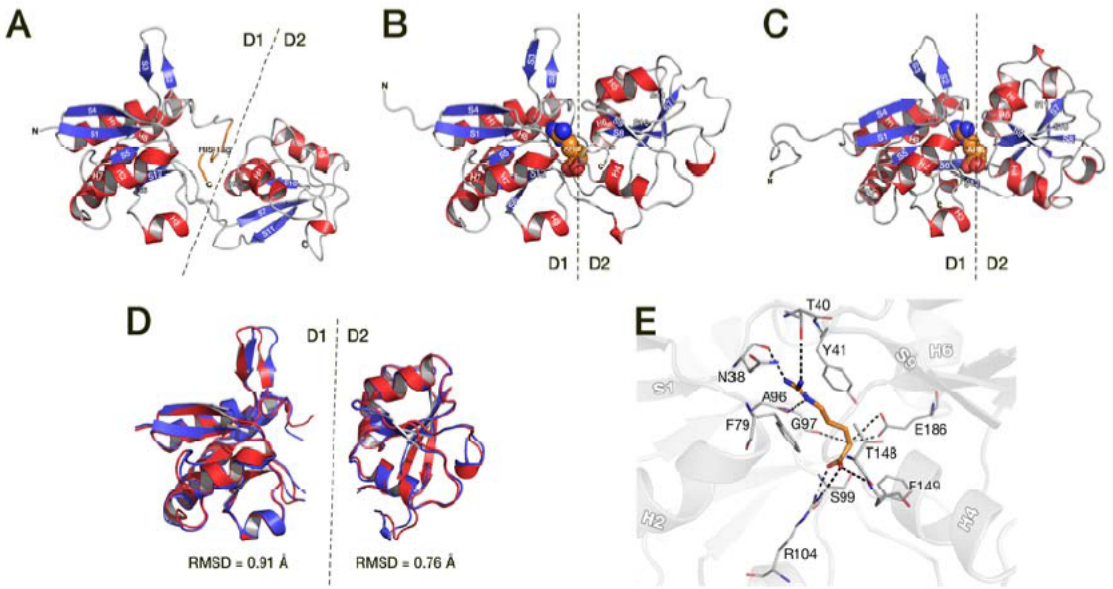


Figure 4.1: *Structures of CT ArtJ (A), CPn ArtJ (B), and G. Stearothermophilus ArtJ part C (pdb-entry 2Q2A). The structures are shown in cartoon representation, with α helices colored red, β strands colored blue and unstructured regions shown as gray tubes. Secondary-structure elements (according to DSSP) are labeled H1 to H8 for helices and S1 to S12 for β strands, following the polypeptide chain from N- to C-terminus. The HIS-tag occupying the binding site in the CT ArtJ structure is labeled and colored orange. The arginine ligand co-crystallized with CPn and G. Stearothermophilus ArtJ is labeled and displayed as van-derWaals spheres (C: orange, N: blue; O: red). Domains D1 and D2 are identified and separated by a dashed line. D) Independent superposition of domains D1 and D2 from CT (red) and CPn (blue) ArtJ. For D1, C-α atoms of residues E34-E120 and V215-N256 (CT ArtJ residue numbering) were used for superposition and calculation of the reported C-α RMSD. For D2, residues I121-K129 and P135-W214 were used. E) Arginine binding site in the crystallographic structure of CPn ArtJ. The arginine ligand and the binding site residues interacting with it are represented with sticks (C: orange for ligand, grey for binding-site residues; N: blue; O: red).*

As expected, the structures of these two proteins are very similar to that of *Geobacillus stearothermophilus* ArtJ, recently solved in its arginine-bound state (PDB code 2Q2A, Figure 4.1C) [73], despite a sequence identity of only 21% between *G. stearothermophilus* and *C. pneumoniae ArtJ*.

Like other periplasmic binding proteins (PBPs), CT and CPn ArtJ adopt a type II PBP fold [74] featuring two $\alpha - \beta$ domains with the arginine-binding site positioned at their interface. The N-C terminal domain (D1) comprises from the N-terminus to residue 120 and from residue 215 to the C-terminus (residue numbering corresponding to CT ArtJ). This domain shares the same structural features in CT and CPn ArtJ.

The central domain (D2) encompasses residues 121 to 214 and displays only

minor structural differences between the two proteins, mainly the presence of two additional β strands along residues 121-124 and 141-144 in CPn ArtJ (residue numbering corresponding to CT ArtJ) (S8 and S9 in Figure 4.1B). This high structural similarity between CPn and CT ArtJ is confirmed by α-carbon root-mean-square differences of 0.91Å(D1) and 0.76Å(D2) (Figure 4.1D). In CPn ArtJ, an extra density was observed in the binding region. This density has been attributed to a bound arginine molecule for the following reasons:

- the density appeared despite the absence of ligand in the model used in refinement
- there is an excellent fit of an Arg molecule in the experimental density that would not fit Lys or His
- its position is identical to the one occupied by Arg in 2Q2A [73](Figure 4.1C)

This aminoacid was captured by the protein during expression and retained during purification despite extensive washes in Ni affinity and a gel filtration step.

The Arg binding site from CPn ArtJ is similar to that described for 2QA2 and is also made of two polar regions flanking a hydrophobic region (Figure 4.1E) [73]. The arginine is held by hydrogen bonds with residues N38, T40, A96, G97, S99, R104, F149, E186 and further enclosed in the binding site by the aromatic side chains of Y41, F79, F149 and by M98, T148. There is no arginine density in CT ArtJ but, in this case, the C-terminal His-tag folded back into the ligand-binding region and a density was observed for all His residues.

4.4 Experimental epitope determination

Mapping of epitopes was performed using both polyclonal sera and monoclonal antibodies. In either case, cross-recognition of the protein from one species by the serum or monoclonal antibodies generated against the other was also tested.

In order to map epitopes recognized by sera from mice immunized with CT and CPn ArtJ, overlapping sequences of 15-residue peptides corresponding to both proteins were synthesized on a cellulose membrane and tested for binding of polyclonal antibodies by immunoblotting. As shown in Figure 4.3 anti-CPn ArtJ polyclonal antiserum recognized a number of different peptides within the N-terminal region of CPn ArtJ, spanning residues 51 to 130. No linear peptides were recognized from residues 131 to 259.

A similar pattern was detected when anti-CPn ArtJ serum was incubated with CT ArtJ spotted membranes. When incubation was performed with anti-CT ArtJ polyclonal antiserum, the positive CT ArtJ peptides were found only within three restricted regions, with one located at the N-terminus and corresponding to residues

16 to 30 and the remaining two located at the C-terminus, and interestingly, the anti-CT ArtJ serum did not strongly recognize any linear peptides of CPn ArtJ.

These results show that linear epitopes are partly common to both variants of ArtJ, but that the immunogenicity of the two antigens is different.

In addition monoclonal antibodies (mAbs) against CT and CPn ArtJ were obtained and used to identify epitope regions by three mapping protocols based on limited antigen proteolysis. MAbs were selected according to their capability to recognize the antigen both by immunoblotting experiments (dot blot and western blotting on recombinant ArtJ) and by FACS analysis on chlamydial EBs (Elementary Bodies). The monoclonal antibodies tested recognized the same peptides overlapping with regions identified by polyclonal antisera and belonging to D1.

4.5 Computational epitope predictions

To include information on ArtJ dynamics in the computational prediction of epitope regions, dynamics of the apo and Arg-bound forms of CT and CPn ArtJ were first studied by MD simulation, based on the crystallographic structures described previously. For each system, five individual simulations of 200 ns were performed in explicit water. The apo forms, as expected for ligand-free PBPs, exhibited large-scale (semi-rigid body) inter-domain movements, whereas the Arg-bound forms were relatively rigid. The structures sampled in these simulations were then clustered by conformation and used for energy-based and electrostatic-desolvation-based predictions of epitope regions.

Epitope predictions were performed by means of MLCE method based on the integration of the topological information available from the atomic-contact matrix with the energetic information attainable through the energy-decomposition approach, allowing the mapping of the principal energy couplings in a protein undergoing dynamics (as described in the previous chapter).

Results of the predictions are reported in Figure 4.2 and identify disctinct epitope maps for the two proteins, supporting a physico-chemical basis for their different antigenic properties.

In particular, the D1 domain of CPn ArtJ shows a higher population of possible antigenic regions than the corresponding domain in CT ArtJ (the predicted epitopes in CT ArtJ D1 are also present in CPn ArtJ D1, the latter having additional ones), while the distribution in the respective D2 domains is more similar. Globally, CPn ArtJ appears to have a higher immunogenic potential, with more predicted epitopes and a larger number of epitope residues.

In addition epitopes on these proteins have been computationally predicted with another technique, which relies on electrostatic-desolvation penalties [75]. This method is based on the idea that similar to other protein-protein interactions, the formation of antibody-antigen complexes requires removal of water molecules from the binding interface region. An approach to calculate the electrostatic energy change upon displacing water from the protein surface was used to systematically investigate the surface desolvation properties of CT and CPn ArtJ. The results of this epitope predictions on ArtJ are summarized in Figure 4.2 and correlate well with the experimental analysis (Figure 4.3).

As with MLCE method, the predictions suggest qualitative differences between the two domains and predict distinct epitope maps for the two proteins, and as a matter of fact the results are remarkably consistent with those of the energy-decomposition-based prediction, despite the different physical properties behind the two methods. In Figure 4.2 is shown the consensus predictions obtained combining the two methodologies, consisting in the strict intersection of the sets of epitopes predicted by the two approaches.

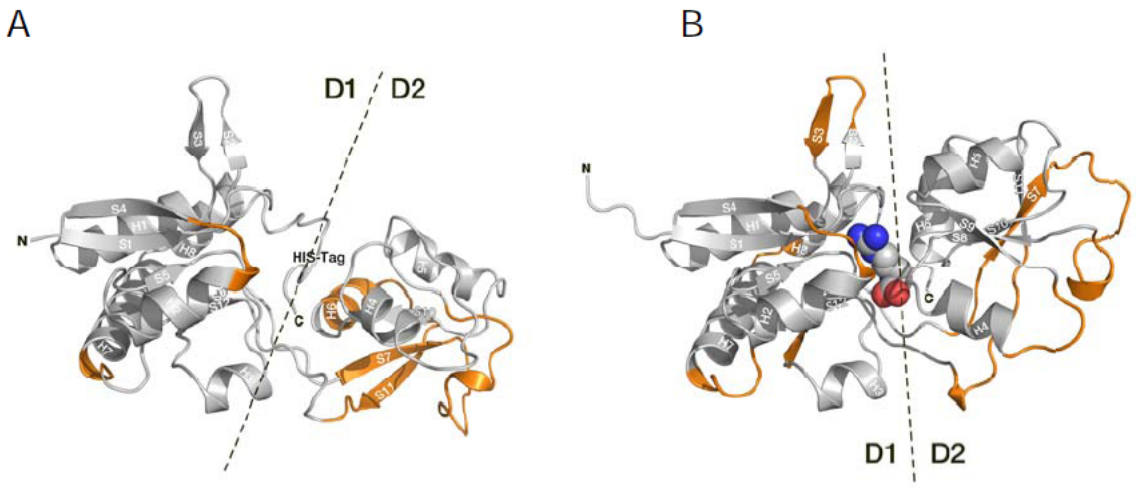


Figure 4.2: *Mapping of consensus predicted epitopes on the X-ray structures of CT (A) and CPn (B) ArtJ. Orange segments correspond to the computationally predicted epitopes. Domains D1 and D2 are identified and separated by a dashed line.*

Furthermore the prediction methods can detect potential differences between the full molecule and the isolated domains, as indicated in Figure 4.3, where some of the epitopes are predicted in the isolated domains and not in the complete X-ray structures. It should be also noted that the prediction of epitopes not detected in the experimental mappings (three with the energy-based method and five with the desolvation-based method) may have a number of causes. Thus, the epitope might still exist but not be detected under the given experimental conditions or the binding-prone

region identified as an epitope might instead participate in other types of interactions, for example, with membrane components.

The consensus prediction shows the same basic features discussed after the independent methods, namely, that despite high sequence and structural similarity the two protein surfaces display local differences in their physico-chemical properties that could potentially have an effect on the relative immunogenic properties of the two molecules. Again, CPn ArtJ appears globally more immunogenic (eight epitopes for four in CT ArtJ), with its D1 domain featuring epitopes that overlap with those in CT ArtJ D1 plus an additional three epitopes.

Predicted epitopes in D2 tend to be less numerous but longer in sequence and larger in surface, arguably indicating participation in less specific interactions (i.e. not with antibodies).

Figure 4.3: *Graphical representation of experimental and computational epitope mappings of CT and CPn ArtJ. Sequence alignment of the two proteins, with conserved residues, is reported in middle row. Beginning and end of D2 are marked with vertical lines. Experimentally determined epitopes are identified by a yellow background. Very low reactivity regions of CPn ArtJ are colored in light blue. Epitope regions predicted by the two computational methods (consensus predicted epitopes) are shown in red font, and highlighted bold when they overlap with experimentally determined epitope regions.*

4.6 Data interpretation

Experimental mapping of antigenic regions in both CPn and CT ArtJ, supported by structure-based computational analysis were crucial in unraveling the antigenic properties of Chlamydial ArtJ. The analysis of the 3D structural organization of the predicted epitope sequences suggests differences between the two proteins at the level of epitope presentation. In particular, both energy-decomposition and electrostatic-desolvation analyses show that putative interaction surfaces in CPn ArtJ are more extensive in

number than in CT ArtJ, suggesting a higher immunogenic character of the former. This is especially so in relation to the D1 domain, which features a larger number of putative epitopes and concentrates a majority of the differences between the epitope sets in the two ortholog proteins, with the CT ArtJ D1 epitopes overlapping with a subset of those predicted for CPn ArtJ. The D2 domain is characterized by a smaller number of predicted epitopes occupying a larger surface, which, combined with the weak immunogenic properties of D2 observed experimentally, could be indicative of a different, less-specific, type of interaction interface.

These results suggest that structural/dynamical differences, determined by relatively small differences in the primary and tertiary arrangements, may define the surface properties underlying differential epitope presentation and recognition by antibodies of the two antigens. It is important to note that the use of high resolution structures, proper characterization of the protein's dynamics and high resolution computational analyses can illuminate those small differences and give access into the structure-dynamics-function relationships at the basis of antibody recognition.

In this context, it is worth underlining that both energy desolvation and MLCE methods predict conformational epitopes, while experimental epitope mapping focuses on linear sequence stretches. This apparent contrast can be solved by noting that the majority of predicted epitopes are made up of aminoacid stretches that are separated along the primary sequence. The latter can thus be conveniently expressed in terms of peptides as in epitope mapping experiments.

In conclusion, the combined analysis of structure, physico-chemical determinants of antibody recognition and epitope mapping allowed the successfully identification and characterization of immunogenic regions in ArtJ. A key aspect of this study was the analysis of two closely related antigens. Indeed, by this comparative and integrated approach, it has been possible to provide a rational for the experimental differences in antibody recognition on the basis of three-dimensional structure and surface properties derived from it, improving knowlewdge in vaccine design.

Chapter 5

Folding and unfolding of small polypeptides

This chapter focuses on the study of the folding and unfolding steps of a 15mer-peptide (called QK) and one of its mutant (called QK_{L10A}), both constituted only by natural aminoacids. In the first part, the biological aspects related to protein and peptide folding are described. Then, in the central section, experimental and computational analyses carried out on QK and QK_{L10A} peptides are explained. In the final part, the data obtained and their relevance for folding process and peptide design are shown. Experimental parts have been performed by Dr. Luca D'Andrea's group at IBB-CNR (Napoli).

This chapter is based on:

- Diana D, Ziaco B, Colombo G, **Scarabelli G**, Romanelli A, Pedone C, Fattorusso R, D'Andrea LD. **Structural determinants of the unusual helix stability of a de novo engineered vascular endothelial growth factor (VEGF) mimicking peptide.** Chemistry. 2008;14(14):4164-6

- Diana D, Ziaco B, **Scarabelli G**, Pedone C, Colombo G, D'Andrea LD, Fattorusso R. **Structural analysis of a helical peptide unfolding pathway.** Chemistry. 2010 May 10;16(18):5400-7

5.1 Protein folding

Protein folding is the physical process by which a polypeptide folds into its characteristic and functional three-dimensional structure from random coil (Figure 5.1). When translated from a sequence of mRNA to a linear chain of aminoacids, a protein exists as an unfolded polypeptide or random coil, which lacks any developed three-dimensional structure. Aminoacids interact with each other to produce a well-defined structural conformation, the folded structure, known as the native state.

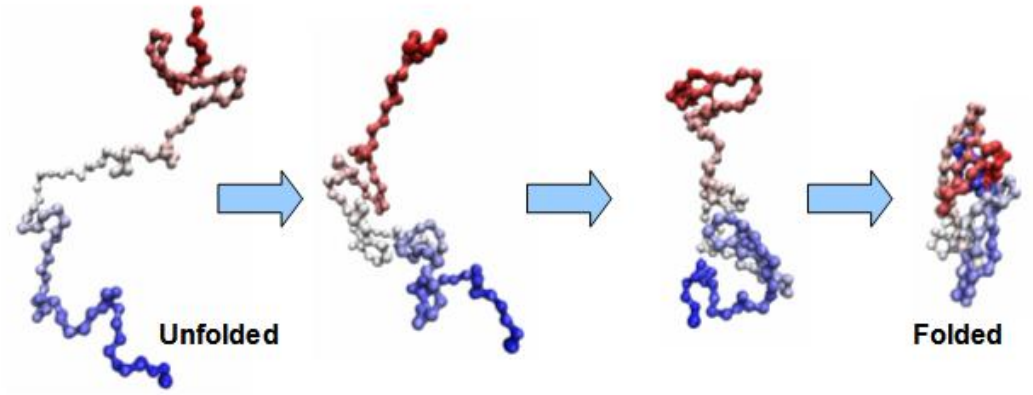


Figure 5.1: *Representation of folding mechanism. Starting from an unstructured one, aminoacid interactions drive the molecule to a folded and functional structure.*

The correct three-dimensional structure is essential to function, although some parts of functional proteins may remain unfolded. In general, failure to fold into the intended shape usually produces inactive proteins with different properties including toxic prions. Several neurodegenerative and other diseases are originated from the accumulation of misfolded (incorrectly folded) proteins in organism tissues.

Importantly, the aminoacidic sequence (or primary structure) of a protein determines its native conformation [76,77] and plays a fundamental role in folding. Proteins can assume spontaneously their functional structure during or after the synthesis. However, the process depends also on other parameters such as solvent type (water or lipid bilayer), salt concentration, temperature and presence of molecular chaperones [78].

Folded proteins usually have a hydrophobic core in which side chain packing has a stabilizing role, while charged or polar side chains occupy the solvent-exposed surface where they interact with surrounding water molecules (Figure 5.2). Minimizing the number of hydrophobic side chains exposed to water is an important driving force behind the folding process [79].

The formation of intramolecular hydrogen bonds provides another important contribution to protein stability. The strength of hydrogen bonds depends on their environment, thus H-bonds enveloped in a hydrophobic core contribute more than H-bonds exposed to the aqueous environment to the stability of the native state [80].

The folding process in vivo often begins co-translationally, so that N-terminus of the protein begins to fold while C-terminal portion is still being synthesized by the ribosome. Specialized proteins called chaperones assist the folding of other proteins [81]. In eukaryotic organisms these molecules are known as heat shock proteins. Although most globular proteins are able to assume their native state unassisted, chaperone-assisted folding is often necessary in the crowded intracellular environment to prevent

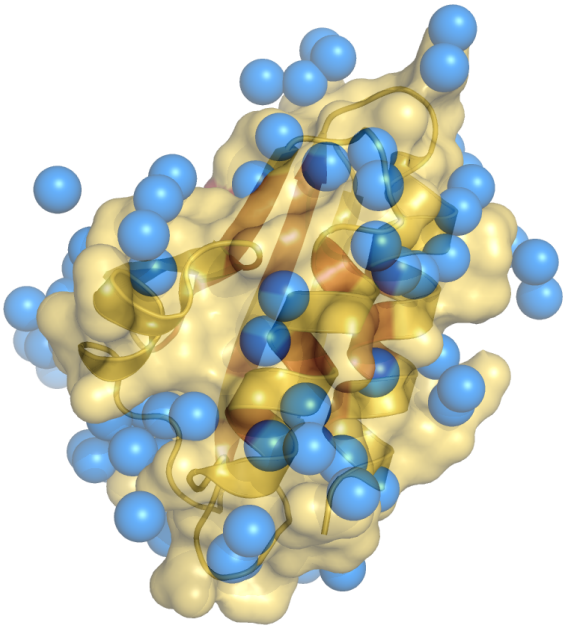


Figure 5.2: *Representation a protein structure. Water molecules surronding the protein are pictured with blue spheres, protein is colored in yellow, red aminoacids are the residues forming the structure core.*

aggregation or misfolding, which are correlated to many pathologies and that may occur as a consequence of exposure to heat or other changes in cellular environment.

On average, studying the complete folding process of a protein using computational simulations is very demanding for actual computer power. In general, the folding into the active conformation, starting from an unstructured one, is carried out with time ranges in the order from nanoseconds to milliseconds, while molecular simulations can cover times of few hundreds of nanoseconds. For this reason, only model systems of these processes have been simulated until now.

In the last year, a new important breakthrough to clarify all the steps of protein folding has been developed by Shaw and coworkes [82]. They designed dedicated CPU cores (called Anton) able to perform MD simulations with an efficiency never reached before. Using this technology, MD folding studies were carried out on two proteins, WW-domain and BPTI, simulating times up to 1 ms (in only three weeks of calculation), highlighting how they gain their stuctured and functional conformations. In particular, in this study simulations revealed that the folding protein followed more or less the same general pattern of movements each time it folded, rather than each folding event having a distinct progression. This fact was something of a surprise because it was not clear from previous modelling and experimental works this would be the case.

In addition, simulating individual proteins for long periods is not the only way to investigate protein folding: networks of computers can also cobble together large

numbers of shorter simulations to explore some key events. Still, this new work sets the stage for extended simulations of dozens, if not hundreds, of other proteins that are less well understood, which could reveal whether all proteins follow a similar set of rules as they fold.

In the end, this new CPU generation has opened remarkably and unthinkable opportunities to characterize biological macromolecule behaviours, allowing the possibility to simulate molecular movements until now inaccessible.

5.1.1 α helix polypeptides

Local formation of secondary structural elements, acting as nucleation sites for the formation of native structures, plays a crucial role in the first phases of protein folding [83]. Thus, an understanding of the folding mechanism, at the highest possible resolution, of synthetic peptides adopting well-defined secondary structures is of great relevance for the comprehension of native protein folding [84]. Peptides that form helices in solution do not show a simple two-state equilibrium between fully folded and fully unfolded structures. Instead, they form a complex mixture of all helix, all coil, and most frequently, central helices with frayed ends [85]. Accordingly, their folding has been often described by using the so-called nucleation-propagation models [86].

These models typically assume a nucleation-growth mechanism with the establishment of a first helical turn representing an entropically unfavorable, slow nucleation reaction, which needs to be balanced by favorable enthalpic contributions. Indeed, according to this mechanism, the nucleation process needs to overcome the largest free-energy barrier, since three residues (corresponding to one helical turn) concomitantly lose their conformational entropy, whereas propagation steps are energetically favorable because of the loss of the conformational entropy of a single residue is balanced by the energy gained from the formation of one extra hydrogen bond [87,88].

This model implies high cooperativity if the nucleation step is rate limiting, since once the first turn has occurred somewhere in the sequence, all subsequent turns would form at essentially the same time. As a matter of fact, previous experiments on the helix-coil transition by employing the laser-induced temperature-jump (Tjump) method have shown that single exponential kinetics, which are characteristic of a two-state system [89,90] can be adequate to describe the helix-coil transition. On the other hand, since the cooperativity of isolated short α helices is weak, they do not generally fold in a two-state fashion, but rather with biexponential kinetics resulting from the coupling of nucleation and the only slightly faster diffusive elongation.

5.2 QK peptide

QK peptide is a designed, α-helical, 15-mer peptide composed only of natural aminoacids (sequence $Ac - KLTWQELYQLKYKGI - NH_2$), which activates VEGF dependent angiogenic response. VEGF is a protein involved in new blood vessel formation and as consequence it is a good target for pathologies like cardiovascular diseases and tumors. Indeed, to live tumor cells require oxygen like any other cell, so the prevention of its assumpion constitutes a way to block tumor cell division.

QK peptide is an interesting molecule to study because it shows an unusual thermal stability. Its design was based on the VEGF binding region to VEGF-receptor (residues 17-25, forming a N-terminal helix), for this reason it shows a high homology degree with it (Figure 5.3). Understanding the aspects related to its stability could have implications in the protein folding field and in the design of helical structured scaffolds, improving the realization of peptides for applications in chemical biology.

As it has been described recently, NMR structure of QK in pure water presents a central helical sequence (residues 4-12), flanked by N- and C-capping regions [91]. QK helical conformation represents an important prerequisite for its biological activity, since the isolated region taken from VEGF protein, corresponding to its binding site, does not assume a helical conformation and does not a have significant biological activity. On the contrary, QK peptide is folded into α helix in solution and it is able to bind its biological target.

Interestingly, QK represents one of the very few examples of bioactive helical designed peptides, composed of only natural aminoacids. To gain an insight into the molecular determinants of QK helical propensity, we examined the effect of the temperature on QK structure through Nuclear Magnetic Resonance (NMR), Circular Dichroism (CD) and Molecular Dynamics (MD) analyses.

```
                15    20    25
VEGF15 Ac-KVKFMDVYQRSYCHP-amide
QK     Ac-KLTWQELYQLKYKGI-amide
```

Figure 5.3: *Sequence alignment of QK peptide with VEGF binding region. The homology between the two segments is at the basis of QK binding to VEGF receptor.*

5.2.1 NMR and CD analyses

Primarily, the aggregation state of the peptide under conditions identical to those used in the NMR structure determination was confirmed by NMR DOSY experiments. DOSY-derived diffusion coefficient value of $1.98 * 10^{-10} m^2 s^{-1}$ is consistent with a QK

monomer state. QK structure variations upon temperature increase were followed by TOCSY experiments. In the 298-343K range only small changes of backbone chemical shifts were observed.

Unusually, chemical shift index (CSI) analysis indicates that at 343K the peptide retains at least the 80% of the helix conformation at 298K and the slight reduction occurs uniformly in 4-12 region.

The thermal behaviour was also analyzed by CD spectroscopy which allowed to extend the temperature range from 278 to 368K. The analysis of the spectra and the dependence of 222nm ellipticity with the temperature, showed that the peptide lose, reversibly, part of its helical structure but neither at 368 K appears to assume a complete random coil conformation.

CD analysis, in accordance with NMR results, indicates that QK retains 79% and 65% of its room temperature helix content at 343 and 368K, respectively, as calculated from the ellipticity at 222nm [92]. Unfortunately, only qualitative structural information could be derived from CD experiments because of the presence in the 9-mer helical segment of three aromatic residues (Trp4, Tyr8, and Tyr12) at positions i, i+4, and i+8, the contribution of which to ellipticity at 222nm is uncertain.

In the end, experimental analyses agreed in the presence of a strong (and unsual for a small polypeptide) helix thermal stability. Up to now, this fact has been reported just for peptides with unnatural constrains [92,93], while QK is formed only by natural aminoacids.

5.2.2 QK peptide MD simulation setup

To assess the determinants of the helix stabilization in solution, experimental structural data have been complemented with extensive all-atom molecular dynamics simulations in explicit water. Five QK structures of the NMR ensemble were used as starting structures, after fitting to NOE constraints. Each model was used as starting conformation for four simulations at the following different temperatures: 300K, 320K, 340K, and 380K. Each 300K MD simulation was 200ns long, while the other ones were each 100ns long. In total 20 MD trajectories were generated for a sampling time of approximately 2.4 microseconds. Moreover, in order to shed light on the folding mechanism, four different simulations with lengths ranging from 50 to 100ns, at 350K, were run from a completely extended structure of the polypeptide.

In each simulation, the isolated peptide structure was first solvated with water in a periodic truncated octahedron, large enough to contain the peptide and 0.9nm of solvent molecules on all sides. The protonation and charge states of the residue sidechains were chosen to be consistent with the solution conditions of the experiments: NH groups were considered with a +1 charge and carboxylic groups were considered to bear a

-1 charge. The system resulted to have a total charge of +2. All solvent molecules within 0.15 nm of any peptide atom were removed. Two Cl^- counterions were added to the system. Different sets of initial velocities obtained from a Maxwellian velocity distribution at the desired temperatures were used to start production runs.

In each case, the system was initially energy minimized with a steepest descent method for 1000 steps. In all simulations the temperature was maintained close to the intended value by weak coupling to an external temperature bath [50], with a coupling constant of 0.1ps.

QK peptide and the rest of the system were coupled separately to the temperature bath. AMBER force field [94] and TIP3P water model [95] were used. LINCS algorithm [49] was used to constrain all bond lengths. A dielectric permittivity, ϵ=1, and a time step of 2fs were used. A cut-off value, set to 0.9nm, was used for the calculation of non-bonded Van der Waals interactions. The calculation of electrostatic forces utilized the PME implementation of the Ewald summation method. In each simulation, the density of the system was adjusted performing the first equilibration runs at NPT condition by weak coupling to a bath of constant pressure (P_0 = 1 bar, coupling time $\tau_P = 0.5ps$) [50].

Production runs have been obtained using NVT conditions, after equilibration, covered the simulation lengths discussed at the beginning of this paragraph. All MD runs and trajectory analyses were performed using GROMACS software package [47]. Peptide configurations were saved every 4ps for subsequent statistical analysis.

Conformational cluster analysis of the trajectories was performed using the method described in Daura et al [15].

5.2.3 QK peptide MD results

All the simulations starting from the helical structures showed a clear, unusual stability of the helix that is maintained even at high temperatures (Figure 5.4), for most of the simulation time, consistently with NMR observations.

To obtain a more global view on peptide behavior, a cluster analysis was performed on a single ensemble obtained from all the structures collected from all the simulations carried out at the same temperature value. Representative structures of the most populated clusters (so the most visited structures in the MD simulations) are represented in Figure 5.5.

Evaluation of structures and stabilizing contacts showed the presence of a network of interactions always involving the hydrophobic side chains of residues 7 and 10 (colored in red in Figure 5.5).

Trajectories were also analyzed in terms of time evolution of root mean square deviation (RMSD) after backbone-backbone superposition, flexibility (RMSF), secondary structure evolution (Figure 5.6) and in terms of stabilizing interactions.

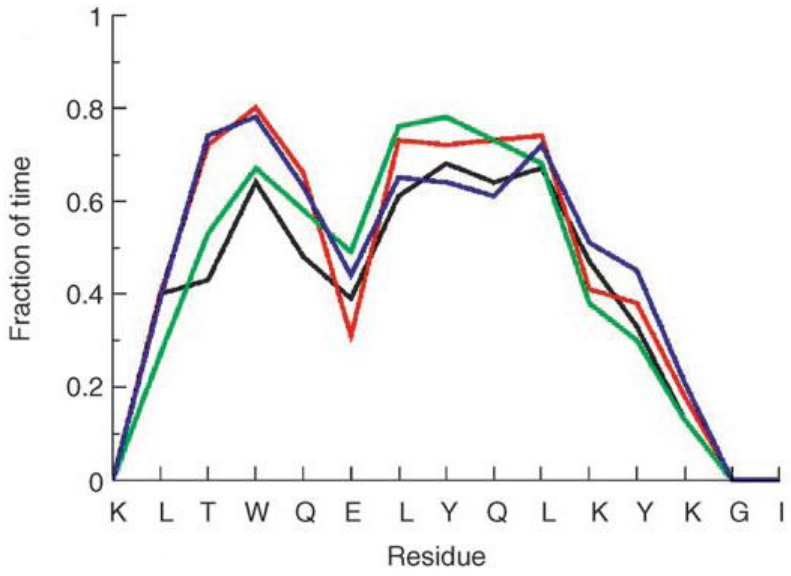


Figure 5.4: *Percentage of time each residue spent in helical conformation at different temperatures. Color code: Black 300K, Red 320K, Green 340K, Blue 380K.*

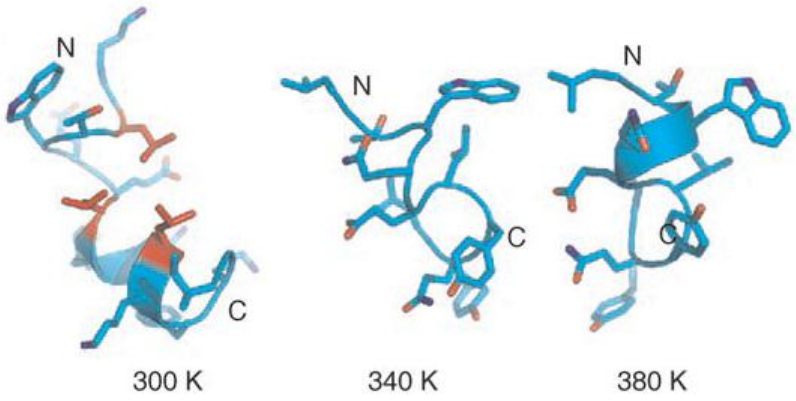
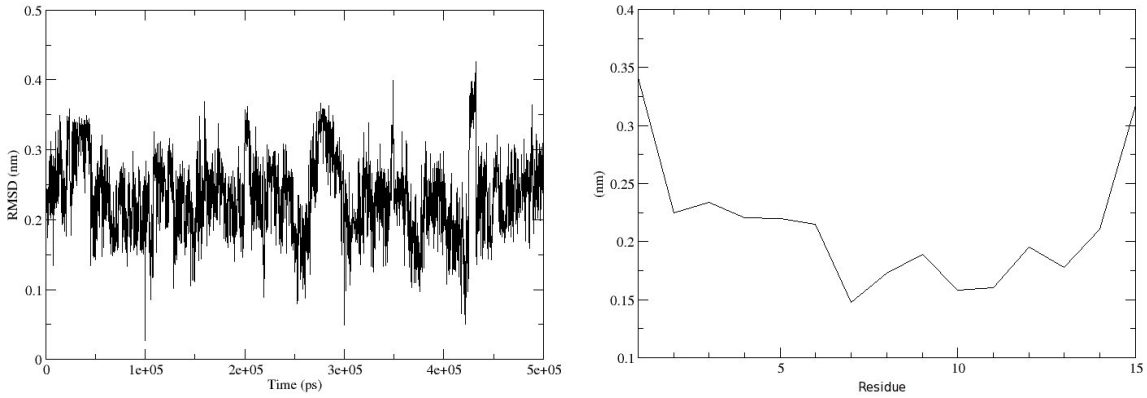


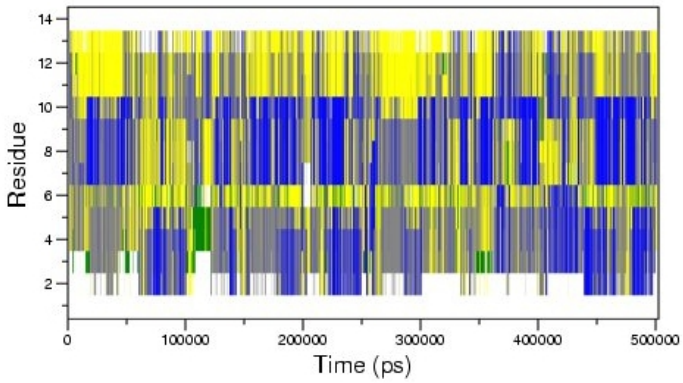
Figure 5.5: *Representative conformations of the main cluster obtained from the analysis of all the trajectories at different temperatures. Leu7 and Leu10 are colored in red in the 300K structure.*

Finally, the first 20ns of each refolding trajectory were analyzed in terms of the percentage of time spent by each residue in a helical conformation, to define the presence of a preferred folding direction.

The analysis of the folding simulations showed a higher tendency for residues located at the N-terminal region to adopt a helical structure in the first events of QK folding (Figure 5.7). This result indicates a relevant role of the N-capping in stabilizing and nucleating the nascent α-helical turn which, then, propagates towards the C-terminal region.



(a) RMSD (b) RMSF



(c) Secondary Structure

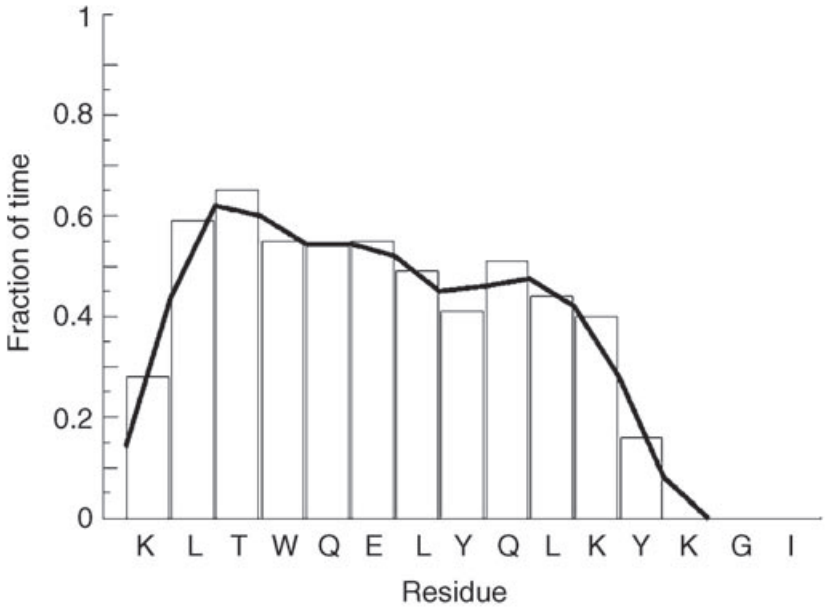
Figure 5.6: *Examples of analyses obtained for the trajectories collected at 320K. (a) RMSD determined using as reference structure the initial helical conformation. (b) RMSF. (c) Secondary Structure during simulation. Color code: Yellow turn, Blue α-helix, Green bend, White coil. All data reported show small variations in peptide conformation. In particular, the central helix formed by residues from 7 to 10 is remarkably stable, highlighting QK stability during all the time simulated. In addition, it is evident the presence of a helix turn at the N terminal (residues 2 to 5). This region is fundamental for the folding process as it forms a structured cap leading the folding of the other peptide residues into helical conformation.*

5.3 QK_{L10A} peptide

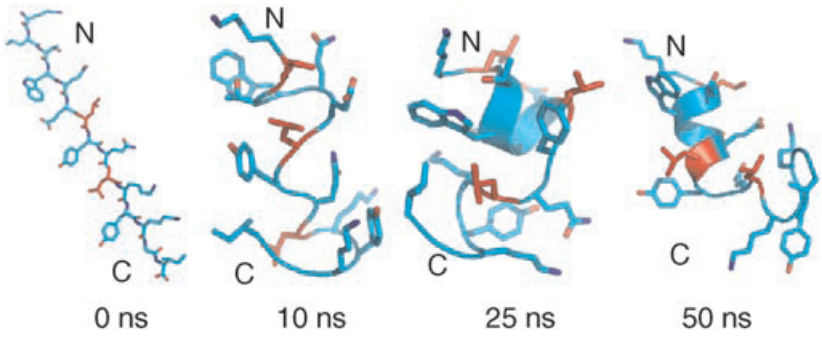
To confirm experimentally these contributions to QK helix stability, we decided to design a novel peptide, called QK_{L10A}, mutating QK sequence: leucine residue in position 10 was replaced by an alanine to perturbe the stabilizing hydrophobic contacts. As the former peptide, this mutant has been analysed with a combination of CD, NMR and MD.

5.3.1 QK_{L10A} CD and NMR results

QK_{L10A} thermal unfolding was investigated by following the variation of the mean residue ellipticity, θ, at 222nm in the 288-353K range. CD analysis is still limited by the



(a) % time spent in α helix



(b) Snapshot refolding structures

Figure 5.7: *(a) Percentage of helical conformation attained by each residue during QK refolding process. (b) Selected structures along the refolding trajectories. Leu7 and Leu10 are highlighted in red.*

presence of three aromatic residues as for QK peptide. Anyway, in this context, peptide helical unfolding appears to be complete above 333K, since no significant variation of the mean residue ellipticity is observed at higher temperatures.

To determine the aggregation state of QK_{L10A}, diffusion-ordered spectroscopy (DOSY) experiments were performed under conditions identical to those used for NMR spectroscopy structure determination. DOSY measurements at 298K provided a diffusion coefficient value of $1.98 * 10^{-10} m^2 s^{-1}$, which is equivalent to that measured for QK peptide and corresponds to what is expected for a monomeric 15-mer helical peptide.

NOESY assignments of QK_{L10A} peptide showed an extensive $H_N - H_N(i, i+1)$, $H_\alpha - H_N(i, i+3)$, $H_\alpha - H_\beta(i, i+3)$ net of cross peaks primarily observed in the 4-12 region of the peptide. This confirmed the high helical propensity of the central region, as already

indicated by the chemical shift index (CSI) analysis based on H_α chemical shifts.

The NMR spectroscopy structure of QK_{L10A} showed a well-defined helix encompassing residues 4-12 in agreement with the CSI data and very similar to what was observed for the QK peptide.

To better investigate the QK_{L10A} thermal unfolding, Dr. D'Andrea and coworkers recorded a set of NMR spectra for the chemical shift proton assignment of QK_{L10A} every 5 Kelvin degrees in the range 288-343 K.

Plots of H_α resonances of QK_{L10A} helical residues as a function of temperature clearly indicate that helical unfolding occurs in the 303-333K range and could be fitted by using unconstrained two- or three-state sigmoidal curves (Figure 5.8).

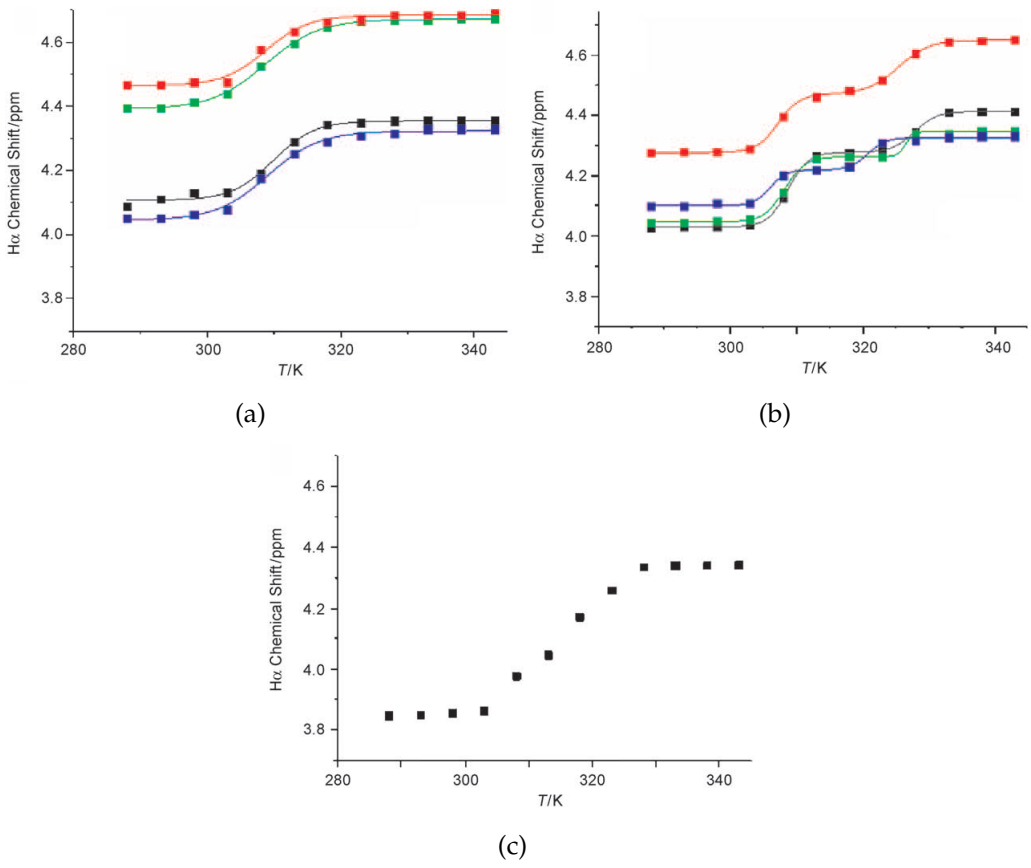
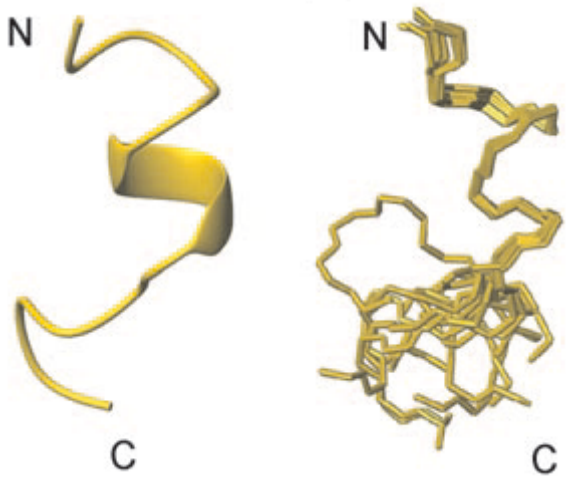


Figure 5.8: *QK_{L10A} H_α chemical shifts plotted versus temperature. (a) Experimental data for residues W4 (green), A10 (black), K11 (blue), and Y12 (red) are fitted with a sigmoidal function. (b) residues E6 (black), L7 (green), Y8 (red), and Q9 (blue) are fitted with a double step function. (c) Q5 could not be fitted with reasonable error.*

In particular, H_α chemical shift curves of residues Trp4, Ala10, Lys11, and Tyr12 (Figure 5.8a) present two-state sigmoidal-like behavior with a melting temperature of about 309K and with the unfolding process concluding at around 320K. On the contrary, Glu6, Leu7, Tyr8, and Gln9 curves show three-state behavior, with two melting temperatures included in the 306-308K and 321-327K ranges (Figure 5.8b). The H_α

chemical shift temperature dependence of Gln5 could not be fitted with reasonable errors, by using either of the two sigmoidal curves, but clearly shows an unfolding process concluding at around 328K (Figure 5.8c).

On the basis of these observations, NMR spectroscopy structure of QK_{L10A} at 313K was determined, in order to derive more resolved conformational preferences when the terminal residues of the helix (residues 4, 5, 10, 11 and 12) appear to be largely unfolded. The NMR structure of QK_{L10A} at 313K (Figure 5.9) shows that the helical conformation is reduced to a single turn, including residues 6 to 9. At the same time, the C-terminal region is largely disordered, whereas the residues at the N-terminal side of the room-temperature helix lost their helical conformation in favor of a distorted turn.



(a) (b)

Figure 5.9: *NMR spectroscopy structures of QK_{L10A} at 313K. (a) Ribbon model of a representative structure and (b) backbone superposition of the 20 energy-minimized structures.*

5.3.2 QK_{L10A} MD simulations results

Equilibrium conformational properties of QK_{L10A} were further investigated by MD simulations in explicit water at different temperatures. The same protocol used for QK peptide has been applied to simulate and analyse QK_{L10A} mutant.

Five different starting structures, corresponding to five different models present in the NMR spectroscopy ensemble, were used for MD simulation. Each model was run for 100ns at four different temperatures of 300, 320, 340, and 380K. A total timespan of 500ns was, thus, run at each temperature, for a total simulation time of 2μs.

Backbone atom-positional root mean square deviation (RMSD) from the α-helical structure determined by NMR spectroscopy at room temperature has been calculated as a function of time (Figure 5.10).

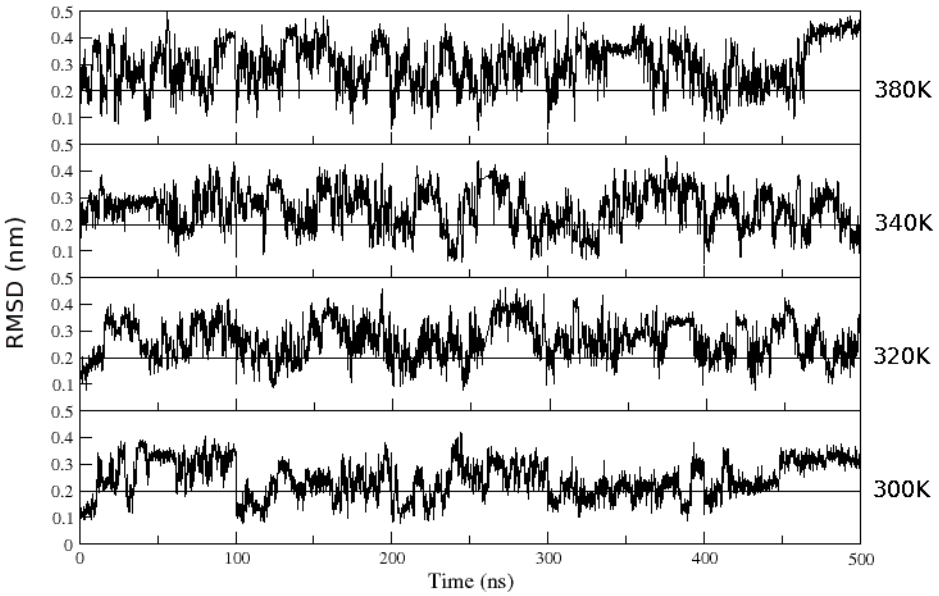


Figure 5.10: *Time evolution during MD simulations of the backbone atom-positional RMSD from the α-helical structure determined at 300K by NMR spectroscopy. The line corresponding to 0.2nm indicates the RMSD threshold with respect to the structure determined by NMR spectroscopy under which a structure is considered folded into a native-like helical conformation.*

At 300K, the structure of QK_{L10A} in water appeared to be stable, with RMSD values consistent with the ideal helical structure (lower than 0.2nm) populated for about 28% of the simulation time. Structural cluster analysis based on an RMSD-similarity criterion proved that, indeed, the three most populated clusters account for mostly α-helical structures.

Increasing the temperature determines a decrease of the ideal helical populations. Notably, a significant amount of the α-helical population is still present at 320 and 340K (20% and 24%, respectively), which is in qualitative agreement with the experimental data. At 380K the peptide is mostly unfolded (the α-helical structure population decreasing to around 14%).

Interestingly, cluster analysis shows that in all cases a significant percentage of residues are still in helical conformation (Figure 5.11). In particular, the central region of QK_{L10A} (residues 6 to 9) is structured as a helical turn in representation of the most populated clusters at 320, 340, and 380K, suggesting that this specific part of the peptide may be the local elementary structure on which full folding is initiated.

The NMR spectroscopic structure of QK_{L10A} at 313K (Figure 5.9b) shows that the helical conformation is reduced to a single turn, including residues 6 to 9.

Backbone atom positional RMSD from this structure was then calculated for each of the temperatures to assess the ability of MD simulations to sample the structures of possible intermediates or partly folded structures (Figure 5.12).

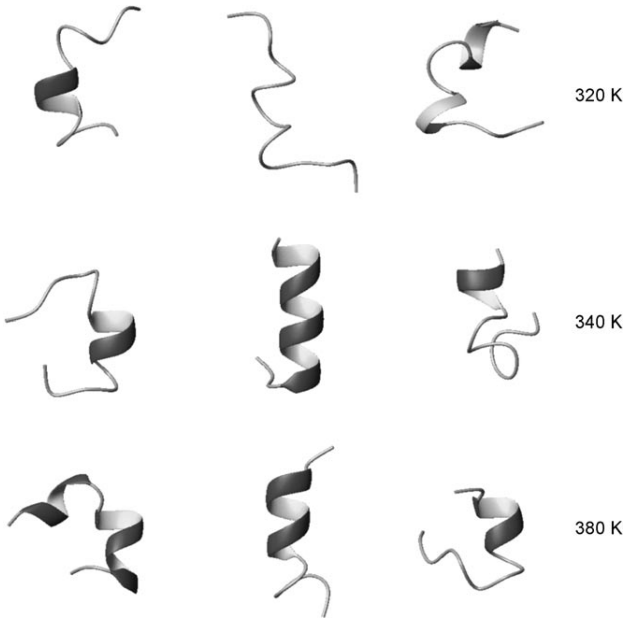


Figure 5.11: *Representative structures of the most populated clusters as derived by MD simulations in explicit water at 320, 340, and 380K.*

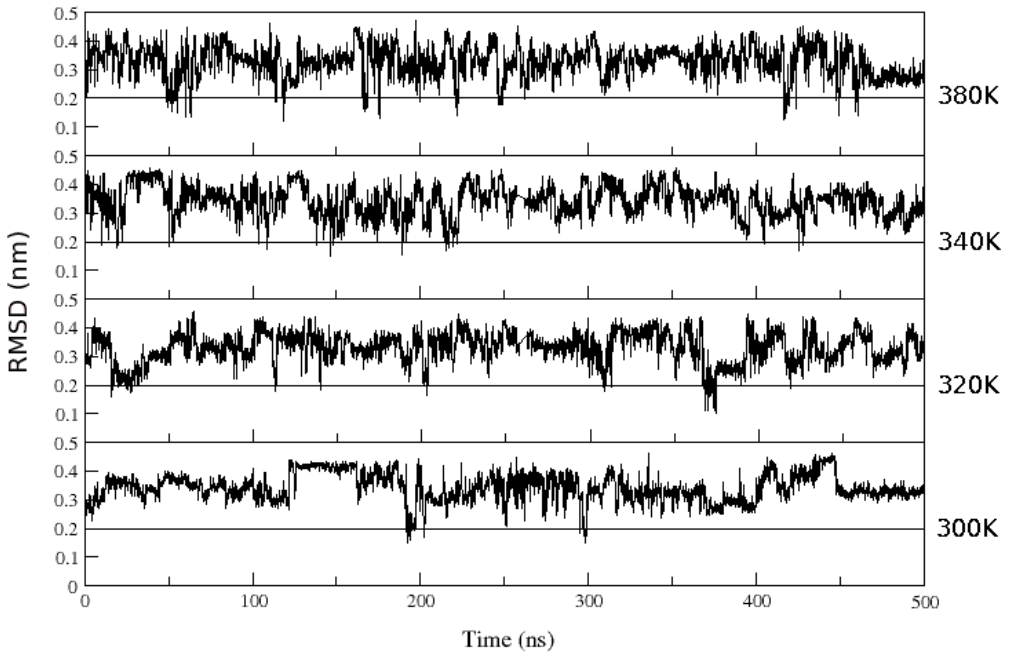


Figure 5.12: *Time evolution during MD simulations of the backbone atom-positional RMSD from the helical-turn-containing structure determined by NMR spectroscopy at 313K. The line corresponding to 0.2nm indicates the RMSD threshold with respect to the structure determined by NMR spectroscopy under which a structure is considered folded into a native-like helical conformation.*

Interestingly, the percentage of structures with low RMSD compared with NMR spectroscopy structure at 313K is more significant at increasing temperatures (320 to

380K). Within the limitations of MD simulations, it is important to note that the highest population of structures similar to the folding intermediate is indeed observed at the simulation temperature (320K) closer to the experimental conditions (313K). In this context, the simulations at different temperatures provide a molecular picture and model for the unfolding curves determined experimentally.

The unfolding evolution of QK_{L10A} was also characterized in terms of percentage of initial helical structure for each residue in the peptide molecules at different temperatures (Figure 5.13). In these cases, only the central residues preserve a high amount of helical secondary structure. In agreement with experimental data, C-terminal residues appear to be the first ones to undergo a transition to a random coil, followed by the N-terminal residues.

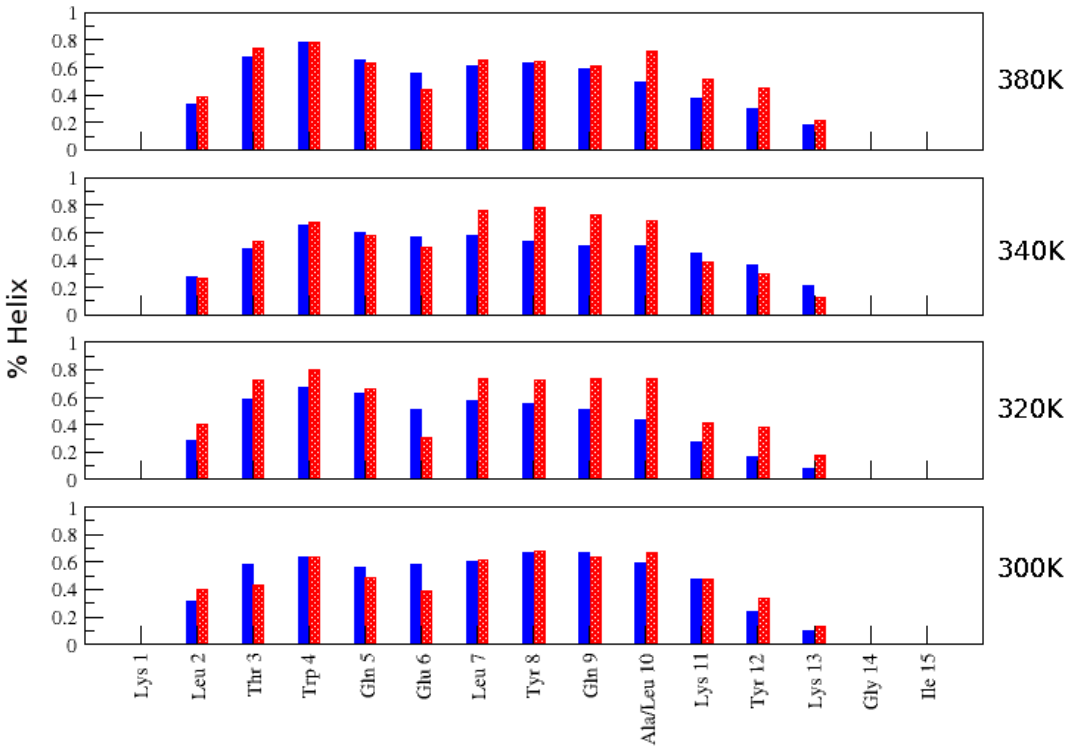


Figure 5.13: *Histogram of the amount of time spent in α-helix per each residue. In blue is represented QK_{L10A} mutant peptide residues, in red QK peptide residues. As it is shown central residues of both peptides are in α-helix conformation for longer time, even though on average in case of QK peptide the % of time is higher.*

Finally, a refolding simulation was run starting from a completely extended conformation, RMSD time evolution with respect to the folded or intermediate structures was monitored (Figure 5.14) in order to characterize the folding steps. Within the 100ns of the simulation, QK_{L10A} is able to fold into a native-like helical conformation. Conformations with low RMSD values with respect to the folding intermediate are sporadically populated in the last 40ns of the refolding simulation. Interestingly, these structures are visited immediately before the peptide visits native-like structures.

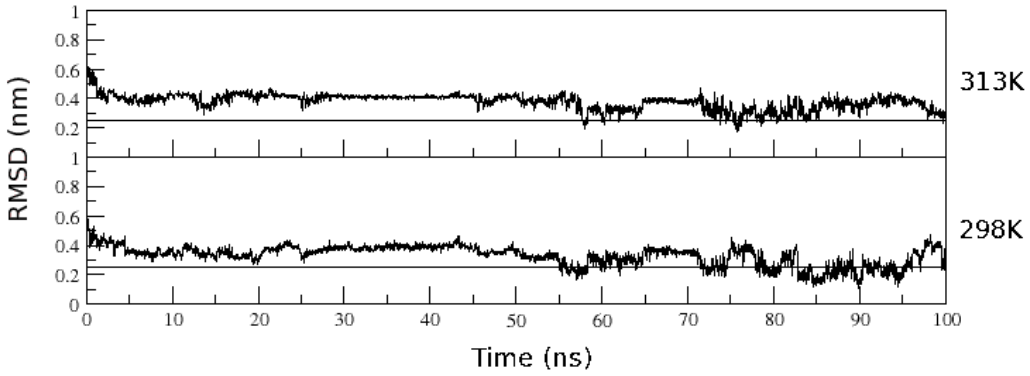


Figure 5.14: *Refolding simulation from a completely extended conformation of the sequence. The lower panel depicts the time evolution of the RMSD of the structures visited during the refolding simulation with respect to the ideal α-helical structure determined by NMR spectroscopy at 298K. The upper panel shows the same quantity calculated with respect to the helical-turn-containing structure determined by NMR spectroscopy at 313K.*

5.4 Discussion

QK is a designed, 15-mer peptide composed only of natural aminoacids, which is a helical in pure water and is able to efficiently activate the VEGF-dependent angiogenic response [91, 96].

To gain a more detailed understanding of the molecular basis of QK helical structure, which appears to be strictly related to its biological activity, we carried out a QK stability characterization. Data collected through experimental analyses (CD and NMR) showed that the QK peptide has an unusual thermal stability up to 368K that prevented us from obtaining a structural depiction of the folding/unfolding pathway of the helix. Nonetheless, using MD computational approach, the structural determinants of QK thermal stability were identified, in particular, two leucine residues (in position 7 and 10) were playing an important role in the strong stability. This fact allowed us to design the less stable QK_{L10A} mutant, which appeared to be largely unfolded at 343K.

Results that provide information about the thermal unfolding process of the QK_{L10A} helix at the molecular level were subsequently obtained, by using a combination of experimental (CD and NMR spectroscopy) and computational (MD) techniques. CD analysis indicates that the helical content in QK_{L10A} is significantly reduced in the 288-333 K temperature range. Moreover, CD unfolding curve shows a broad transition, as previously observed for the helix-coil transition of peptides with similar helical content, which is compatible with the nucleation-propagation folding-unfolding model [85].

Nonetheless, the absence of a plateau at the lowest values of the CD temperature dependency plot has required a more detailed molecular analysis of QK_{L10A} thermal unfolding obtained through NMR spectroscopic techniques. QK_{L10A} structure, obtained through NMR spectroscopy, is in complete agreement with the secondary structure

indicated by the H_α based CSI.

The solution structure of QK_{L10A} is almost identical to that of QK and consists of a helix spanning the residues from Trp4 to Tyr12 and of two disordered N- and C-terminal tails. Although being significantly less populated than in the QK peptide, as shown by CD spectra and CSI analyses, the helical conformation in QK_{L10A} involves exactly the same central region as in QK. MD simulations indicate that at 300K QK_{L10A} populates an ideal helical conformation for 28% of the simulation time and partial α-helical structures for some of the remaining time.

To obtain higher resolution structural details of the thermal unfolding pathway of the mutant peptide, we used the temperature dependency of QK_{L10A} proton chemical shifts. In particular, H_α proton chemical shifts have proven to be the most sensitive to follow QK_{L10A} helix unfolding. H_α curves of the 15 residues (Figure 5.8) show that the 4-12 helical structure is almost completely folded at 298K, even though residues 4, 11, and 12 still weakly increase the global helical content in the 288-298K range.

At higher temperatures, helix unfolding initially involves, principally, the C-terminal end (residues 10, 11, and 12) and also, to a lesser extent, the N-terminal end (residue 4). The melting temperatures of this first part of the unfolding process are all between 308 and 309K and those residues conclude their transition to a random coil conformation by 323K.

The central residues 5-9, during this first phase of the peptide unfolding, reduce their helical content, but do not completely unfold, indicating that during the unwinding of the terminal ends there is also a reduction of the helicity in the central turn. These central aminoacids experience a second and definitive unfolding process, for which the melting temperature is in the 321-327K range and that is completed around 333K.

To obtain higher resolution details of the whole unfolding mechanism, we determined the solution structure of QK_{L10A} at 313K, the highest temperature allowing the collection of enough NOEs for structure determination. On the basis of H_α temperature dependence and $\Delta\delta H_\alpha$ plots (Figure 5.8), at this temperature the unfolding of terminal ends is expected to be largely complete, whereas central residues should still be helical conformation.

NMR spectroscopy structure of QK_{L10A} at 313K (Figure 5.9) is based on 13 $^3J_{HNH_\alpha}$ coupling constants and 83 meaningful NOEs that were mostly observed in the 3-9 residue region, allowing, therefore, a reasonable resolution of this region of the peptide. Comparison of structures at 298 and 313K clearly shows that QK_{L10A} at 313K retains a single turn of helix encompassing residues 6 to 9, whereas N-terminal tail folds in a distorted turn and whole 10-15 residue region is mostly disordered.

Interestingly, MD analyses of the most populated clusters at higher temperatures support the view that QK_{L10A} structures preserve the central helix turn. Moreover, simulation data also are in agreement with the contention that the first part of the structure to undergo thermal unfolding is the C-terminal tail, followed by the N terminus.

Finally, MD analyses indicated that QK_{L10A} is able to refold from an extended structure in 100ns, which is a typical time range reported for helix folding. The central helix conformations are more frequently visited immediately before the peptide fully forms its helical structure, therefore they are possible key intermediates for QK_{L10A} helix folding. These data suggest that the most representative folding-unfolding pathways of QK_{L10A} may actually involve the helical turn 6-9 as intermediate and necessary step to drive the collapse of the full sequence to the native ensemble.

QK_{L10A} peptide folding behaviour is different from QK, for which the formation of a helix turn at the N-terminal was the first step leading peptide folding. The reason at the basis of this fact is constituted by the absence of the leucine residue in position 10. As an alanine has a side chain smaller than a leucine, the residue in position 10 in QK_{L10A} is not able to be part of the hydrophobic interaction network, which is instead present in QK structure. As consequence there is not a complete formation of this contact network and because of this in QK_{L10A} peptide is not present the formation of a stable helix turn at N-terminal region.

To conclude, the rational, MD-based design of QK_{L10A} allowed us to turn the original QK peptide, for which the accessible conformational space is limited only to helical structures at different temperatures, into a molecule endowed with substantially higher conformational freedom (QK_{L10A} peptide). This, in turn, made possibile to carry out experimental and theoretical characterization at atomic level of intermediate structures that survive at temperatures slightly higher than 300K.

Both experimental and theoretical investigations detected a stable helical structure spanning residues 6 to 9. This organized substructure defined a local elementary motif that may act as a nucleus necessary to drive the subsequent collapse of the whole peptide to the full helical structure.

It is important to underline that this is one of the first cases in which a real local elementary structure is observed to be formed at an atomic resolution and stable in a short peptide, at different temperatures by a consensus of different techniques.

Computational and experimental data presented herein allows us to draw a folding-unfolding picture for the small peptide QK_{L10A}, composed only of natural aminoacids, compatible with a nucleation-propagation model. The central turn including residues 6 to 9 represents the most probable site for helix nucleation, which then propagates toward the two terminal ends. The N terminus could already be prearranged in a distorted turn, whereas the helix appears to propagate through a mostly disordered C-terminal backbone.

This mechanism also provides a reasonable explanation for the extra stability of QK compared with that of QK_{L10A}. The side chain-side chain interaction between Leu7 and Leu10, which the MD analyses already showed to be important for QK stability, could play a crucial role in further stabilizing and enlarging the central helix nucleating turn.

Leu10 mutation to Ala reduces the stability and the length of the central turn, therefore strongly diminishing the global QK_{L10A} thermal stability.

This study, besides contributing to the basic field of peptide helix folding, is useful to gain an insight, in general, into the design of stable helical peptides, which could find applications as molecular scaffolds upon which to graft functional residues to modulate protein-protein interactions. In particular, it provides guidelines that are helpful to modify the chemical and physical properties of the QK molecule, one of the few peptides with pro-angiogenic activity in vivo.

Chapter 6

Conclusion and final remarks

Proteins are organic compounds made of aminoacids organized in a linear chain and folded into a globular form. They can perform a variety of tasks and participate in virtually every process within cells.

Many proteins are enzymes that catalyze biochemical reactions and are vital to metabolism, other proteins can have structural or mechanical functions, such as proteins in the cytoskeleton forming a system of scaffolding that maintains the cell shape. Proteins can play important roles in cell signaling, immune response, cell adhesion, cell cycle, etc.

For all these reasons, it is worth to say that the study and the analysis of their characteristics has a great relevance to understand biological life. Both experimental and theoretical approaches can be combined to carry out general and complete studies to obtain information on protein characteristics and functions.

In my thesis work I used Molecular Dynamics (MD) simulations applied to protein and peptide structures to analyse their microscopic properties.

The first part of this thesis was focused on the characterization of antigenic proteins. In particular, a new method for epitope prediction (MLCE, Matrix of Local Coupling Energies) has been developed in order to improve vaccine design process.

MLCE is based on energetic and structural information contained in a protein structure. Through the calculation of non-bonded energetic interactions between each pair of residues and taking into account only close residues in the protein architecture, MLCE is able to identify with good approximation epitope regions, spanning from loops (which are the most common type of structure for epitopes) to ordered segments like α helices and β sheets. Importantly, MLCE method is not knowledge based and does not require the use of any training set of known antigens.

Furthermore, to make MLCE available to the scientific community, a public web server (BEPPE, Binding Epitope Prediction from Protein Energetics) has been implemented with it (web-link: http://158.109.215.216/upload.php?UserName= 103). Only a

protein structure file (in pdb format) is required to use BEPPE and the output is sent to an email address specified by the user.

BEPPE performance has been compared with other public web-servers able to predict epitope locations. Results showed that BEPPE reached a better precision value than all the others. In particular, on a set of known antigens, we saw that almost the half of the residues identified by BEPPE has been found to be bound specifically by antibodies.

Finally, as the innovative approach developed carries out predictions considering only physical-chemical properties of each residue in a protein structure, it has the potential to predict general protein interaction sites. In this context, we are currently analyzing a number of protein complexes to evaluate the performance on these cases.

The second part of the thesis was related to the characterization of folding/unfolding processes of small α helical polypeptides, homologue to VEGF protein binding region and showing a biological activity in solution. In particular, we have studied two molecules, called QK and QK_{L10A}, constituted by only 15 natural aminoacids. They show the same sequence except for position 10, QK presents a leucine, while QK_{L10A} has an alanine. Several MD simulations have been carried out on both peptides using different temperatures (ranging from 300 to 380K) in order to characterize the folding/unfolding steps of their structures.

In the case of QK we were unable to completely unfold it. Analyses allowed to identify the residues responsible of the unsual peptide stability and how it folds into the α helical shape. In special, the interaction between leucine residues in position 7 and 10 was importantly contributing to the thermal stability. As confirmation of this fact QK_{L10A} peptide was less stable at high temperatures.

In addition, we saw that the peptides have diverse folding pathways. In the case of QK, N-terminal region forms a helix turn which is then followed by the rest of the peptide, while for QK_{L10A} only the central region folds into a α helix.

All these data were in agreement with experimental results, obtained with CD and NMR, carried out on these two peptides by Dr. D'Andrea's group at IBB-CNR institute.

To conclude, we applied and developed new theoretical approaches, based on Molecular Dynamics simulations, on two interesting but different biological cases: prediction of epitopes on antigenic proteins and small peptide folding. Our final aim was to obtain useful information on these two fields in order to integrate, support and drive experimental studies. In both cases, we were able to gain relevant insight through the implementation of classical and new analyses techniques whose application can be transferred also to other systems, demostrating the importance of the combination of theoretical and experimental methodologies in scientific research.

Bibliography

[1] Rinaudo C.D., Telford J.L., Rappuoli R., Seib K.L. Vaccinology in the genome era. J Clin Invest. **2009** Sep;119(9):2515-25.

[2] Tiana G., Simona F., De Mori G. M., Broglia R. A., Colombo G. Understanding the determinants of stability and folding of small globular proteins from their energetics. Protein Sci. **2004** Jan;13(1):113-24.

[3] Ragona L., Colombo G., Catalano M., Molinari H. Determinants of protein stability and folding: comparative analysis of beta-lactoglobulins and liver basic fatty acid binding protein. Proteins. **2005** Nov 1;61(2):366-76.

[4] Colacino S., Tiana G., Colombo G. Similar folds with different stabilization mechanisms: the cases of Prion and Doppel proteins. BMC Struct Biol. **2006** Jul 21;6:17.

[5] Ponomarenko J., Bui H. H., Li W., Fusseder N., Bourne P. E., Sette A., Peters B. ElliPro: a new structure-based tool for the prediction of antibody epitopes. BMC Bioinformatics. **2008** Dec 2;9:514.

[6] Altschul S. F., Gish W., Miller W., Myers E. W., Lipman D. J. Basic local alignment search tool. J Mol Biol.**1990** Oct 5;215(3):403-10.

[7] Amela I., Cedano J., Querol E. Pathogen proteins eliciting antibodies do not share epitopes with host proteins: a bioinformatics approach. PLoS One. **2007** Jun 6;2(6):e512.

[8] Selkoe D.J. Folding proteins in fatal ways. Nature **2003** 426:900-904.

[9] Bartlett A. I., Radford S. E. An expanding arsenal of experimental methods yields an explosion of insights into protein folding mechanisms Nat Struct & Mol Bio **2009** 16, 582-588.

[10] Huang C. Y., Klemke J. W., Getahun G., DeGrado W. F., Gai F. Helix formation via conformation diffusion search **2002** Proc. Natl. Acad. Sci. USA 99 2788-2793.

[11] Clark R. J., Craik D. J. Native chemical ligation applied to the synthesis and bioengineering of circular peptides and proteins. Biopolymers. **2010**;94(4):414-22.

[12] Meli M., Colombo G. Molecular simulations of peptides: a useful tool for the development of new drugs and for the study of molecular recognition. Methods Mol Biol. **2009**;570:77-153.

[13] Darden, T., York, D., Pedersen, L. Particle mesh Ewald: An N-log(N) method for Ewald sums in large systems. J. Chem. Phys. **1993**, 98, (10089-10092).

[14] Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., Pedersen, L. G. A smooth particle mesh Ewald potential. J. Chem. Phys. **1995**, 103, 8577-8592.

[15] Daura X., Gademann K., Jaun B., Seebach D., Gunsteren W. F. V., Mark A. E. Peptide folding: when simulation meets experiment. Angew. Chemie Intl. Ed. **1999**, 38, 236-240.

[16] Kabsch W., Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen bonding and geometrical feature. Biopolymers **1983**,22,2576-2637.

[17] Tama F. Normal mode analysis with simplified models to investigate the global dynamics of biological systems. **2003** Protein Pept. Lett., 10, 119-132.

[18] Tirion M. M. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. **1996** Phys. Rev. Lett., 77, 1905-1908.

[19] Colacino S., Tiana G., Broglia R. A., Colombo G. The determinants of stability in the human prion protein: insights into folding and misfolding from the analysis of the change in the stabilization energy distribution in different conditions. Proteins. **2006** Mar 15;62(3):698-707.

[20] Morra G., Baragli C., Colombo G. Selecting sequences that fold into a defined 3D structure: A new approach for protein design based on molecular dynamics and energetics. Biophys Chem. **2010** Feb;146(2-3):76-84. Epub 2009 Oct 31.

[21] Sharp K. A., Honig, B. Electrostatic interactions in macromolecules: Theory and applications. Annu. Rev. Biophys. Biophys. Chem. **1990** 19, 301-332.

[22] Davis M. E.,McCammon J. A. Electrostatics in biomolecular structure and dynamics. Chem. Rev. **1990** 90, 509-521.

[23] Borchardt J. K. The History of Bacterial Meningitis Treatment Drug. News Perspect. **2004**, 17, 219-24.

[24] Cartwright K., Noah N., Peltola H. Meningococcal disease in Europe: epidemiology, mortality, and prevention with conjugate vaccines. Report of a European advisory board meeting Vienna, Austria, 6-8 October, 2000. Vaccine **2001** Aug 14;19(31):4347-56.

[25] Gotschlich E. C., Liu T. H., Artenstein M. S. Preparation and immunochemical properties of the group A, group B, and group C meningococcal polysaccharides J. Exp. Med.**1969** 129:1349-1365.

[26] Hayrinen J., Jennings H., Raff H. V., Rougon G., Hanai N., Gerardy-Schahn R., Finne J. J. Infect. Dis. Antibodies to polysialic acid and its N-propyl derivative: binding properties and interaction with human embryonal brain glycopeptides. **1995**, 171, 1481-90.

[27] Jodar L., Feavers I.M., Salisbury D., Granoff D.M. Development of vaccines against meningococcal disease. Lancet **2002**, 359, 1499-508.

[28] Seib K. L., Rappuoli R., Difficulties in developing neisserial vaccines. In: C.A. Genco and L.M. Wetzler, Editors, Neisseria: molecular mechanisms of pathogenesis, Horizon Scientific Press, Norwich, UK **2010**, pp. 195-226.

[29] Tettelin H., Saunders N. J., Heidelberg J., Jeffries A. C., Nelson K. E., Eisen J. A., Ketchum K. A., Hood D. W., Peden J. F., Dodson R. J., Nelson W. C., Gwinn M. L., DeBoy R., Peterson J. D., Hickey E. K., Haft D. H., Salzberg S. L., White O., Fleischmann R. D., Dougherty B. A., Mason T. , Ciecko A., Parksey D. S., Blair E., Cittone H., Clark E. B., Cotton M. D., Utterback T. R., Khouri H., Qin H., Vamathevan J., Gill J., Scarlato V., Masignani V., Pizza M., Grandi G., Sun L., Smith H. O., Fraser C. M., Moxon E. R., Rappuoli R., Venter J. C. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. Science. 2000 Mar 10;287(5459):1809-15.

[30] Pizza M., Scarlato V., Masignani V., Giuliani M. M., Aricò B., Comanducci M., Jennings G. T., Baldi L., Bartolini E., Capecchi B., Galeotti C. L., Luzzi E., Manetti R., Marchetti E., Mora M., Nuti S., Ratti G., Santini L., Savino S., Scarselli M., Storni E., Zuo P., Broeker M., Hundt E., Knapp B., Blair E., Mason T., Tettelin H., Hood D. W., Jeffries A. C., Saunders N. J., Granoff D. M., Venter J. C., Moxon E. R., Grandi G., Rappuoli R. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. Science. **2000** Mar 10;287(5459):1816:20.

[31] Rappuoli R. Reverse vaccinology. Curr Opin Microbiol. **2000** Oct;3(5):445-50.

[32] Giuliani, M. M., Adu-Bobie J., Comanducci M., Aricò B., Savino S., Santini L., Brunelli B., Bambini S., Biolchi A., Capecchi B., Cartocci E., Ciucchi L., Di Marcello F., Ferlicca F., Galli B., Luzzi E., Masignani V., Serruto D., Veggi D., Contorni M., Morandi M., Bartalesi A., Cinotti V., Mannucci D., Titta F., Ovidi E., Welsch J. A., Granoff D., Rappuoli R., Pizza M. A universal vaccine for serogroup B meningococcus. Proc. Natl. Acad. Sci. U. S. A. **2006** 103:10834-10839.

[33] Rappuoli R. **2008**. The application of reverse vaccinology, Novartis MenB vaccine developed by design [abstract]. Presented at the 16th International Pathogenic Neisseria Conference. Rotterdam, The Netherlands. http://neisseria.org/ipnc/2008/.

[34] Rappuoli R, Covacci A. Reverse vaccinology and genomics. Science. **2003** Oct 24;302(5645):602.

[35] Dormitzer P. R., Ulmer J. B., Rappuoli R. Structure-based antigen design: a strategy for next generation vaccines. Trends Biotechnol. **2008** 26:659-667.

[36] van Regenmortel M. H. V. Mapping Epitope Structure and Activity: From One-Dimensional Prediction to Four-Dimensional Description of Antigenic Specificity **1996** A Companion to Methods in Enzymology 9, 465-472.

[37] Rubinstein N. D., Mayrose I., Halperin D., Yekutieli D., Gershoni J. M., Pupko T. Computational characterization of B-cell epitopes. Mol Immunol. **2008** Jul;45(12):3477-89. Epub 2007 Nov 26.

[38] Ponomarenko J. V., Bourne P. E. Antibody-protein interactions: benchmark datasets and prediction tools evaluation. BMC Struct. Biol. **2007** 7:64.

[39] Ma B. Y., Wolfson H. J., Nussinov R. Protein functional epitopes: hot spots, dynamics and combinatorial libraries. Curr. Opin. Struct. Biol. **2001** 11:364-369.

[40] Novotny J., Handschumacher M., Haber E., Bruccoleri R. E., Carlson W. B., Fanning D. W., Smith J. A., Rose G. D. Antigenic determinants in proteins coincide with surface regions accessible to large probes (antibody domains). Proc Natl Acad Sci U S A. **1986** Jan;83(2):226-30.

[41] Zhou T., Xu L., Dey B., Hessell A. J., Van Ryk D., Xiang S. H., Yang X., Zhang M. Y., Zwick M. B., Arthos J., Burton D. R., Dimitrov D. S., Sodroski J., Wyatt R., Nabel G. J., Kwong P. D. Structural definition of a conserved neutralization epitope on HIV-1 gp120. Nature. **2007** Feb 15;445(7129):732-7.

[42] Bizebard T., Gigant B., Rigolet P., Rasmussen B., Diat O., Bosecke P., Wharton S. A., Skehel J. J., Knossow M. Structure of influenza virus haemagglutinin complexed with a neutralizing antibody. Nature. **1995** Jul 6;376(6535):92-4.

[43] Morra G., Colombo G. Relationship between energy distribution and fold stability: insights from molecular dynamics simulations of native and mutant proteins Proteins. **2008** Aug;72(2):660-72.

[44] Ferreiro D. U., Hegler J. A., Komives E. A., Wolynes P. G. Localizing frustration in native proteins and protein assemblies. Proc Natl Acad Sci U S A. **2007** Dec 11;104(50):19819-24. Epub 2007 Dec 5.

[45] Wang W., Lim W. A., Jakalian A., Wang J., Wang J., Luo R., Bayly C. I., Kollman P. A. An analysis of the interactions between the Sem-5 SH3 domain and its ligands using molecular dynamics, free energy calculations, and sequence analysis. J Am Chem Soc. **2001** May 2;123(17):3986-94.

[46] Berendsen H. J. C., Grigera J. R., Straatsma P. R. The missing term in effective pair potentials. J. Phys. Chem. **1987** 91:6269-6271.

[47] 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. **2004** Gromacs User Manual version 3.2. www.gromacs.org.

[48] Scott W. R. P., Hunenberger P. H., Tironi I. G., Mark A. E., Billeter S. R., Fennen J., Torda A. E., Huber T., Kruger P., Gunsteren W. F. V. The GROMOS biomolecular simulation program package. J.Phys.Chem.A **1999** 103:3596-3607.

[49] Hess B., Bekker H., Fraaije J. G. E. M., Berendsen H. J. C. A linear constraint solver for molecular simulations. J.Comp.Chem. **1997** 18:1463-1472.

[50] Berendsen H. J. C., Postma J. P. M., van Gunsteren W. F., Di Nola A., Haak J. R. Molecular dynamics with coupling to an external bath. J. Chem. Phys. **1984** 81:3684-3690.

[51] Haste Andersen P., Nielsen M., Lund O. Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. Protein Sci. **2006** Nov;15(11):2558-67. Epub 2006 Sep 25.

[52] Fawcett T. An introduction to ROC analysis. Pattern Recognit. Lett. **2006** 27:861-974.

[53] de Vries S. J., Bonvin A. M. J. J. How proteins get in touch: interface prediction in the study of biomolecular complexes. Curr. Protein Pept. Sci. **2008** 9:394-406.

[54] Chavali G. B., Papageorgiou A. C., Olson K. A., Fett J. W., Hu G., Shapiro R., Acharya K. R. The crystal structure of human angiogenin in complex with an antitumor neutralizing antibody. Structure. **2003** 11:875-885.

[55] Hubbard S. J., Campbell S. F., Thornton J. M. Molecular recognition. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. J.Mol.Biol. **1991** 220,507-530.

[56] Suhre K., Sanejouand Y. H. ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. Nucleic Acids Res. **2004** Jul 1;32.(Web Server issue):W610-4.

[57] Haste Andersen P., Nielsen M., Lund O. Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. Protein Sci. **2006** Nov;15(11):2558-67. Epub 2006 Sep 25.

[58] Sweredoski M. J., Baldi P. PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure. Bioinformatics. **2008** Jun 15;24(12):1459-60. Epub 2008 Apr 28.

[59] Bloom J. D., Labthavikul S. T., Otey C. R., Arnold F.H. Protein stability promotes evolvability. Proc Natl Acad Sci U S A. **2006** Apr 11;103(15):5869-74. Epub 2006 Mar 31.

[60] Baker D. A surprising simplicity to protein folding. Nature **2000** 405:39-42.

[61] Grantcharova V. P., Riddle D. S., Santiago J. V., Baker D. Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain Nat Struct Biol. **1998** Aug;5(8):714-20.

[62] Denisova G. F., Denisov D. A., Yeung J., Loeb M. B., Diamond M. S., Bramson J. L. A novel computer algorithm improves antibody epitope prediction using affinity-selected mimotopes: a case study using monoclonal antibodies against the West Nile virus E protein. Mol Immunol. **2008** Nov; 46(1):125-34. Epub 2008 Aug 29.

[63] Lange O. F., Lakomek N. A., Fares C., Schroder G. F., Walter K. F., Becker S., Meiler J., Grubmuller H., Griesinger C., de Groot B.L. Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. Science. **2008** Jun 13;320(5882):1471-5.

[64] Ho B. K., Agard D. A. Probing the flexibility of large conformational changes in protein structures through local perturbations. **[2009** PLOS Comput. Biol. 5:e1000343.

[65] Sweredoski M. J.,Baldi P. COBEpro: a novel system for predicting continuous B-cell epitopes. Protein Eng. Des. Sel. **2009** 22:113-120.

[66] Friedrichs M. S., Eastman P., Vaidyanathan V., Houston M., Legrand S., Beberg A. L., Ensign D. L., Bruns C. M., Pande V. S. Accelerating molecular dynamic simulation on graphics processing units. J Comput Chem. **2009** Apr 30;30(6):864-72.

[67] Klepeis J. L., Lindorff-Larsen K., Dror R. O., Shaw D. E. Long-timescale molecular dynamics simulations of protein structure and function. Curr Opin Struct Biol. **2009** Apr;19(2):120-7. Epub 2009 Apr 8.

[68] Halling-Brown M. D., Moss D. S., Sansom C. E., Shepherd A. J. A computational Grid framework for immunological applications. Philos Transact A Math Phys Eng Sci. **2009** Jul 13;367(1898):2705-16.

[69] Zugel U., Kaufmann S. H. Role of heat shock proteins in protection from and pathogenesis of infectious diseases. Clin Microbiol Rev. **1999** Jan;12(1):19-39.

[70] Finco O., Bonci A., Agnusdei M., Scarselli M., Petracca R., Norais N., Ferrari G., Garaguso I., Donati M., Sambri V., Cevenini R., Ratti G., Grandi G. Identification of new potential vaccine candidates against Chlamydia pneumoniae by multiple screenings. Vaccine. **2005** Jan 19;23(9):1178-88.

[71] Makarova K. S., Mironov A. A., Gelfand M. S. Conservation of the binding site for the arginine repressor in all bacterial lineages. Genome Biol. **2001**;2(4):RESEARCH0013. Epub 2001 Mar 22.

[72] Sali A., Blundell T. L. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol **1993** 234, 779-815

[73] Vahedi-Faridi A., Eckey V., Scheffel F., Alings C., Landmesser H., Schneider E., Saenger W. Crystal structures and mutational analysis of the arginine-, lysine-, histidine-binding protein ArtJ from *Geobacillus stearothermophilus*. Implications for interactions of ArtJ with its cognate ATP-binding cassette transporter, Art(MP)2. J Mol Biol. **2008** Jan 11;375(2):448-59.

[74] Fukami-Kobayashi K., Tateno Y., Nishikawa K. Domain dislocation: a change of core structure in periplasmic binding proteins in their evolutionary history. J Mol Biol. **1999** Feb 12;286(1):279-90.

[75] Fiorucci S., Zacharias M. Prediction of protein-protein interaction sites using electrostatic desolvation profiles. Biophys J. **2010** May 19;98(9):1921-30.

[76] Anfinsen C. B. Principles that Govern the Folding of Protein Chains Science. **1973** 181 (96): 223-230.

[77] Berg J. M., Tymoczko J. L., Stryer L. Protein Structure and Function. Biochemistry. **2003** San Francisco: W. H. Freeman.

[78] van den Berg B., Wain R., Dobson C. M., Ellis R. J. Macromolecular crowding perturbs protein refolding kinetics: implications for folding inside the cell. EMBO J. **2000** 19 (15): 3870-5

[79] Pace C., Shirley B., McNutt M., Gajiwala K. Forces contributing to the conformational stability of proteins. FASEB J. **1996** 10 (1): 75-83.

[80] Deechongkit S., Nguyen H., Dawson P. E., Gruebele M., Kelly J. W. Context Dependent Contributions of Backbone H-Bonding to β-Sheet Folding Energetics Nature **2004** 403 (45): 101-5.

[81] Lee S., Tsai F. Molecular chaperones in protein quality control J. Biochem. Mol. Biol. **2005** 38 (3): 259-65.

[82] Shaw D. E., Maragakis P., Lindorff-Larsen K., Piana S., Dror R. O., Eastwood M. P., Bank J. A., Jumper J. M., Salmon J. K., Shan Y., Wriggers W. Atomic-level characterization of the structural dynamics of proteins. Science. **2010** Oct 15;330(6002):341-6.

[83] Daggett V., Fersht A. The present view of the mechanism of protein folding. Nat Rev Mol Cell Biol. **2003** Jun;4(6):497-502.

[84] Bartlett A. I, Radford S. E. An expanding arsenal of experimental methods yields an explosion of insights into protein folding mechanisms. Nat Struct Mol Biol. **2009** Jun;16(6):582-8.

[85] Doig A. J. Recent advances in helix-coil theory. Biophys Chem. **2002** Dec 10;101-102:281-93.

[86] Streicher W. W., Makhatadze G. I. Calorimetric evidence for a two-state unfolding of the beta-hairpin peptide trpzip4. J Am Chem Soc. **2006** Jan 11;128(1):30-1.

[87] Werner J . H., Dyer R. B., Fesinmeyer R. M., Andersen N. H. Dynamics of the Primary Processes of Protein Folding: Helix Nucleation J. Phys. Chem. B **2002**, 106, 487-494.

[88] Ihalainen J. A., Paoli B., Muff S., Backus E. H., Bredenbeck J., Woolley G. A., Caflisch A., Hamm P. Alpha-Helix folding in the presence of structural constraints. Proc Natl Acad Sci U S A. **2008** Jul 15;105(28):9588-93.

[89] Lednev I. K., Karnoup A. S., Sparrow M. C., Asher S. A. Nanosecond UV Resonance Raman Examination of Initial Steps in α-Helix Secondary Structure Evolution J. Am. Chem. Soc. **1999**, 121, 4076-4077.

[90] Lednev I. K., Karnoup A. S., Sparrow M. C., Asher S. A. α-Helix Peptide Folding and Unfolding Activation Barriers: A Nanosecond UV Resonance Raman Study J. Am. Chem. Soc. **1999**, 121, 8074-8086.

[91] D'Andrea L. D., Iaccarino G., Fattorusso R., Sorriento D., Carannante C., Capasso D., Trimarco B., Pedone C. Targeting angiogenesis: structural characterization and biological properties of a de novo engineered VEGF mimicking peptide. Proc Natl Acad Sci U S A. **2005** Oct 4;102(40): 14215-20.

[92] Shepherd N. E. , Hoang H. N., Abbenante G., Fairlie D. P. Single turn peptide alpha helices with exceptional stability in water. J Am Chem Soc. **2005** Mar 9;127(9):2974-83.

[93] Wang D., Chen K., Kulp I J. L., Arora P. S. Evaluation of Biologically Relevant Short α-Helices Stabilized by a Main-Chain Hydrogen-Bond Surrogate J. Am. Chem. Soc., **2006**, 128 (28), pp 9248-9256.

[94] Cornell W. D., Cieplak P., Bayly C. I., Gould R. I., Merz K. M. Jr., Ferguson D. M., Spellmeyer D. C., Fox T., Caldwell J. W., Kollman P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules J. Am. Chem. Soc., **1995**, 117 (19), pp 5179-5197.

[95] Jorgensen W. L., Chandrasekhar J., Madura J., Impey R. W., Klein M. L. Comparison of simple potential functions for simulating liquid water J. Chem.Phys. (1983) 79; 926-935.

[96] Santulli G., Ciccarelli M., Palumbo G., Campanile A., Galasso G., Ziaco B., Altobelli G. G., Cimini V., Piscione F., D'Andrea L. D., Pedone C., Trimarco B., Iaccarino G. In vivo properties of the proangiogenic peptide QK. J Transl Med. **2009** Jun 8;7:41.

Acknowledgments

First of all, I would like to thank Giorgio for giving me the possibility to take the PhD course in Industrial Chemistry and for his support during all these years. Giorgio has a huge comprehensive scientific knowledge and he is a very good person, I am very pleased to have had the chance to work with him.

Actually, in these years I tried to teach him something, but I failed. In particular, my main concern is that he still believes Inter FC won 4 italian championships (the paperboard one does not even deserve to be considered!) and 1 UEFA champions league in a honest way. Unfortunately, he has not accepted the truth yet... ☺

However, apart from jokes (written in the first four lines!☺), I really enjoyed to be part of such a good team formed by excellent people, who also have had a relevant role in the improvement of my abilities, so thanks to (in order of appearance) Giulia, Max, Elisabetta, Marco, Ale e Rubben.

What's more, I would like to say that, in this group, I gained the title of "best student" in Doctor Torella's class, which is a quite remarkable result! ☺

In addition, I owe a special thank to all the CRACU's guys for all the lunchtimes spent together!

Furthermore, I would like to thank Dr. Laura Belvisi and Prof. Anna Bernardi for their support in thesis writing and for all the presentations I have done during the course.

Finally, I thank you for reading this thesis!